CVENET: A COMPLEMENTARY VIEW AND EXPERT NETWORK FOR TIME SERIES FORECASTING WITH EXOGENOUS VARIABLES

Anonymous authors
Paper under double-blind review

ABSTRACT

Time series forecasting is essential in many domains. Beyond endogenous variables (i.e., target variables), exogenous variables (i.e., covariates) provide additional predictive information over the history, and often over the forecast horizon as well. However, existing methods for *t*ime *s*eries *f*orecasting with e*x*ogenous variables (TSF-X) mainly focus on how endogenous and exogenous variables interact, and fix in advance how the series are processed before and during this interaction, leading to two shortcomings: 1) they handle non-stationarity with a single treatment, either removing the window statistics or keeping them, although the preferable one may shift along the forecast horizon, and 2) they apply the same functional form to every token, as even mixture-of-experts forecasters mostly route tokens among architecturally identical experts. To address these shortcomings, we propose **CVENet**, a **C**omplementary **V**iew and **E**xpert **Net**work for time series forecasting with exogenous variables. Specifically, Complementary View Construction first builds two multi-scale views of the history, with and without the window statistics removed, each carried by its own stack of encoder layers with dual-correlation attention. In every encoder layer, Token-Wise Complementary Experts then route each token within a frequency-domain and a time-domain expert family and mix the two with a token-dependent gate. Finally, Horizon-Wise View Reconciliation combines the two view forecasts with a weight learned for each forecast step. Extensive experiments on twelve real-world datasets and a tokamak dataset demonstrate that CVENet outperforms state-of-the-art methods.

1 INTRODUCTION

Time series forecasting plays a vital role in many domains, including economics (Huang et al., 2022), traffic (Lv et al., 2014), energy (Alvarez et al., 2011), nuclear fusion (Kates-Harbeck et al., 2019; Wan et al., 2021), and AIOps (Pan et al., 2023). The most direct approach forecasts the target series (i.e., endogenous variables) from its own history, yet in real-world applications, covariate series (i.e., exogenous variables) are often closely related to the target: day-ahead electricity price forecasting rests on forecasts of system load and available generation (Lago et al., 2021), and the power delivered by a wind farm follows the weather around it. Forecasting with exogenous variables incorporates such series as auxiliary inputs and has become an important forecasting paradigm (Wang et al., 2024; Zhou et al., 2025; Qiu et al., 2026).

From a temporal perspective, exogenous variables are of two types: historical exogenous variables, observed over the input window, and future exogenous variables, known over the forecast horizon because they are fixed in advance, as with calendars and scheduled control plans, or can be supplied as forecasts, as with weather. Methods differ in how they admit this information: TiDE (Das et al., 2023) and TFT (Lim et al., 2021) consume future exogenous variables directly, TimeXer (Wang et al., 2024) and CrossLinear (Zhou et al., 2025) work from history alone, and dual-correlation designs jointly model temporal and channel correlations (Qiu et al., 2026; Li et al., 2026b). Generally, using historical and future exogenous variables together exploits the most complete information available, which motivates us to target this setting, while also accommodating scenarios where future exogenous variables are unavailable. However, while existing methods mainly focus on how

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.

endogenous and exogenous variables interact, they fix in advance how the series are processed before and during this interaction, leading to two shortcomings.

First, they handle non-stationarity with a single prescribed treatment. Series in these domains drift in level and scale over time (Liu et al., 2025a). A standard remedy removes the window statistics before modeling and restores them afterwards (Kim et al., 2022; Nie et al., 2023), aligning windows so that shared patterns can be learned, but it also removes what the statistics carry: the level from which the future departs is restored rather than learned, and information useful for prediction can be lost (Liu et al., 2022; Li et al., 2026a). Keeping them preserves this information but leaves windows misaligned, and neither treatment is preferable in general: normalization is disabled for individual datasets (Lin et al., 2024), and under a common backbone, applying it or not reverses the ranking of two decomposition rules (Yu et al., 2024). Recent work balances the two treatments inside attention, by dependency range in TimeBridge (Liu et al., 2025a) or by instance in SeesawNet (Li et al., 2026a). The balance may also shift along the forecast horizon, as the input window tends to be a more reliable reference for the first steps ahead than for later ones. We therefore retain both treatments in a multi-scale decomposition front end, where the level of the window resides in the trend components (Wu et al., 2021; Zeng et al., 2023), and weigh them at each forecast step (Fig. 1a).

Second, they apply the same functional form to every token. Time series superpose periodic and non-periodic structures that call for different inductive biases, yet the perceptrons usually applied to every token handle such differing structures poorly, periodicity in particular (Dong et al., 2025). Mixture-of-experts (MoE), which activates only a few experts for each input, has become a common strategy for scaling and specializing time series models (Shazeer et al., 2017; Shi et al., 2025; Qiu et al., 2025b). However, the expert pools are usually homogeneous: in TFPS (Sun et al., 2025), for example, every expert is the same two-layer perceptron, so routing chooses among parameter instances rather than functional forms. Recent designs introduce heterogeneous experts (Ortigossa et al., 2026; Liu et al., 2025b), but fix the expert types in advance, so routing still selects among experts of a single type. We instead route two families of complementary functional forms, a frequency-domain family with a periodic inductive bias and a time-domain family of perceptrons, alongside an always-on shared expert (Dai et al., 2024), and let the content of each token decide how the two are mixed (Figure 1b).

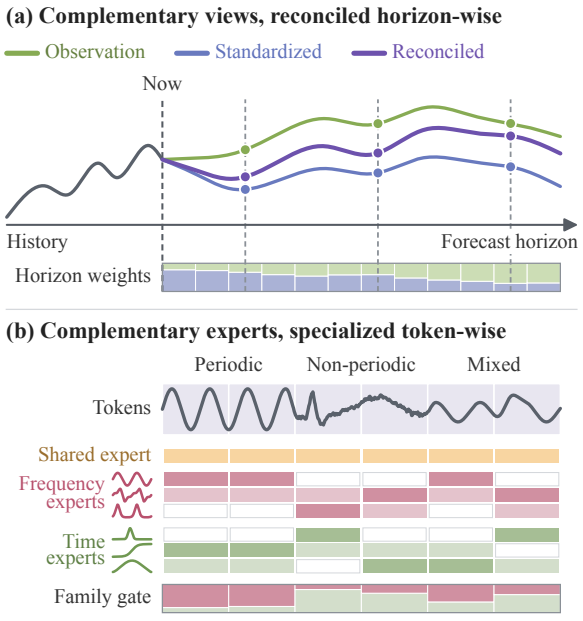


Figure 1: Illustration of complementary views and experts: (a) two view forecasts combined with a weight learned for each forecast step; (b) tokens routed within a frequency-domain and a time-domain expert family and mixed by a token-dependent gate.

To address these shortcomings, we propose **CVENet**, a **C**omplementary **V**iew and **E**xpert **Net**work for time series forecasting with exogenous variables. Specifically, **for the first challenge**, we first introduce *Complementary View Construction*, which builds an observation view and a standardized view of the history at multiple scales, each carried by its own stack of encoder layers with dual-correlation attention (Qiu et al., 2026). Next, **for the second challenge**, inside every encoder layer, *Token-Wise Complementary Experts* mix a frequency-domain and a time-domain expert family for each token. Finally, *Horizon-Wise View Reconciliation* combines the two view forecasts with a weight learned for each forecast step.

Our contributions are summarized as follows:

- We propose CVENet, a general framework for time series forecasting with exogenous variables, which handles the non-stationarity of the history with complementary statistical views and adapts the functional form of the transformation to each token.
- We design Complementary View Construction, which retains both treatments of non-stationarity as two multi-scale views of the history, and Horizon-Wise View Reconciliation, which combines the two view forecasts with a weight learned for each forecast step.
- We design Token-Wise Complementary Experts, which route each token within a frequency-domain expert family with a periodic inductive bias and a time-domain family of perceptrons, and let a token-dependent gate mix the two.
- We evaluate CVENet on twelve real-world datasets under both exogenous regimes, and on EAST-POINT127, a tokamak dataset in which actuator signals drive the electron density. CVENet outperforms state-of-the-art methods.

2 Related Work

2.1 Time Series Forecasting with Exogenous Variables

Classical approaches such as ARIMAX (Williams, 2001) and SARIMAX (Vagropoulos et al., 2016) introduce exogenous drivers through linear regression terms, which limits the modeling of nonlinear and time-varying effects. Deep forecasters differ mainly in how endogenous and exogenous information interact. Early-fusion designs such as TiDE (Das et al., 2023) and NBEATSx (Olivares et al., 2023) integrate the two sources before prediction; TFT (Lim et al., 2021), TimeXer (Wang et al., 2024) and ExoTST (Tayal et al., 2024) make the interaction selective or asymmetric through variable selection or cross-attention; and CrossLinear (Zhou et al., 2025), Aura (Lin et al., 2026) and dual-correlation designs model the dependency structure itself, with DAG (Qiu et al., 2026) encoding temporal and channel correlations in parallel branches and GCGNet (Li et al., 2026b) aligning their joint structure in graph form. Across this progression, the design effort is concentrated on how the two sources interact, while the non-stationarity of each series is handled before that interaction by a single prescribed treatment.

2.2 Non-stationary Time Series Forecasting

Real-world time series are often non-stationary (Kim et al., 2022; Liu et al., 2025a). One line of work normalizes the input, as in RevIN (Kim et al., 2022), Dish-TS (Fan et al., 2023) and SAN (Liu et al., 2023), and another separates the slowly varying trend, as in Autoformer (Wu et al., 2021), FEDformer (Zhou et al., 2022), DLinear (Zeng et al., 2023), xPatch (Stitsyuk & Choi, 2025), Leddam (Yu et al., 2024) and AMD (Hu et al., 2025). The two lines overlap, since normalization removes the level that a decomposition would otherwise absorb into its trend, and this overlap has been used to explain why normalization yields different gains across backbones (Stitsyuk & Choi, 2025). Removing non-stationarity also has a cost, as over-stationarized inputs lose information useful for prediction (Liu et al., 2022). Recent methods therefore balance elimination and retention: TimeBridge (Liu et al., 2025a) removes non-stationarity for short-term dependencies and preserves it for long-term cointegration, and SeesawNet (Li et al., 2026a) fuses attention on normalized and raw sequences with an instance-adaptive gate. These balances are struck inside attention, per dependency or per instance, whereas how the preferable treatment may vary along the forecast horizon has been little explored.

2.3 Mixture-of-Experts for Time Series Forecasting

Mixture-of-experts (MoE) combines networks through a learned gate (Jacobs et al., 1991), and sparse gating activates only a few experts per input (Shazeer et al., 2017). Time series models use homogeneous sparse experts to scale forecasting foundation models (Shi et al., 2025) and to dispatch inputs by latent-distribution clustering (Qiu et al., 2025b) or learned subspaces (Sun et al., 2025); specialization there comes from the data each expert receives, while the experts share one architecture and differ only in their weights. Recent work instead varies the pool itself, introducing heterogeneous experts (Liu et al., 2025b; Ortigossa et al., 2026) or expanding a pool of functionally

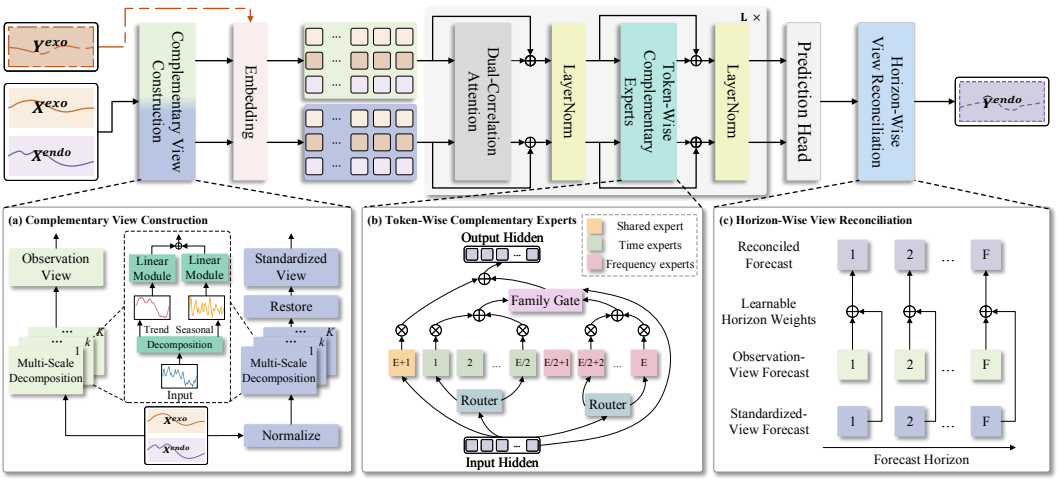


Figure 2: The architecture of CVENet. (a) *Complementary View Construction*, which builds an observation view and a standardized view. (b) *Token-Wise Complementary Experts*, which route each token within a frequency-domain and a time-domain expert family and mix the two by a token-dependent gate. (c) *Horizon-Wise View Reconciliation*, which combines the two view forecasts step by step.

diverse experts when drift is detected (Zhu et al., 2026). Our design routes within each of two expert families and learns a token-dependent mixture between them.

3 Methodology

3.1 Problem Formulation

In the context of exogenous-aware time series forecasting, the historical endogenous time series $X^{\text{endo}} \in \mathbb{R}^{N\times T}$ is accompanied by historical exogenous variables $X^{\text{exo}} \in \mathbb{R}^{D\times T}$ and future exogenous variables $Y^{\text{exo}} \in \mathbb{R}^{D\times F}$, where N is the number of endogenous variables, D is the number of exogenous variables, T is the number of historical time steps, and F is the forecasting horizon. The objective is to forecast the future endogenous series $\hat{Y}^{\text{endo}} \in \mathbb{R}^{N\times F}$ over the horizon F based on the given historical endogenous and exogenous series, which can be formally expressed as follows.

$$\hat{Y}^{\text{endo}} = \mathcal{F}_\theta\left(X^{\text{endo}}, X^{\text{exo}}, Y^{\text{exo}}\right) \tag{1}$$

3.2 Architecture Overview

Figure 2 shows the architecture of CVENet. At the input, we introduce the *Complementary View Construction*, which builds two multi-scale views of the history that differ in how the window statistics are treated. We then carry each view through its own stack of encoder layers with dual-correlation attention (Qiu et al., 2026), and in every encoder layer we apply the *Token-Wise Complementary Experts* to tokens that have already absorbed their correlated context. Each view therefore produces a forecast of its own, and we finally combine the two by *Horizon-Wise View Reconciliation*, with a weight learned for each forecast step.

3.3 Complementary View Construction

Decomposition hands each component of a series to its own learnable map, and since the trend components carry the level of the window, it is also where the non-stationarity of the history is kept or removed. We leave open both the scale of the separation and whether the window statistics are removed.

Let $X = [X^{\text{endo}}; X^{\text{exo}}] \in \mathbb{R}^{C\times T}$ with $C = N + D$ denote the historical series of all variables, endogenous and exogenous alike. Following the series decomposition block of Autoformer (Wu

et al., 2021), a moving average of an input $U \in \mathbb{R}^{C\times T}$ yields its trend component and the remainder yields its seasonal component. We apply this at K scales (Zhou et al., 2022),

$$U_{\mathrm{t}}^{(k)} = \mathrm{AvgPool}_{s_k}(\mathrm{Padding}(U)), \quad U_{\mathrm{s}}^{(k)} = U - U_{\mathrm{t}}^{(k)}, \tag{2}$$

where $s_1 < \cdots < s_K$ are odd kernel sizes and $\mathrm{Padding}$ replicates the two endpoints so that every component keeps the length T. Each component is transformed by its own linear map along the time axis, and the results are added together (Zeng et al., 2023),

$$\mathcal{D}(U) = \sum_{k=1}^{K} \left(U_{\mathrm{t}}^{(k)} W_{\mathrm{t}}^{(k)} + U_{\mathrm{s}}^{(k)} W_{\mathrm{s}}^{(k)} \right) + b, \tag{3}$$

with $W_{\mathrm{t}}^{(k)}, W_{\mathrm{s}}^{(k)} \in \mathbb{R}^{T\times T}$ and $b \in \mathbb{R}^{T}$ shared across variables, so that $\mathcal{D}(U)$ retains the shape of its input.

The observation view applies this operation to the series as it is observed, whereas the standardized view removes the window statistics beforehand and restores them afterwards, following instance normalization without learnable affine parameters (Kim et al., 2022):

$$H^{\mathrm{obs}} = \mathcal{D}_{\mathrm{obs}}(X), \quad H^{\mathrm{std}} = \sigma \odot \mathcal{D}_{\mathrm{std}}((X - \mu) \oslash \sigma) + \mu, \tag{4}$$

where $\mu, \sigma \in \mathbb{R}^{C}$ are the per-variable mean and standard deviation over the window, broadcast over time, with a small constant ϵ added to the variance, $\odot$ and $\oslash$ denote elementwise product and division, and $\mathcal{D}_{\mathrm{std}}$ has the same form as $\mathcal{D}_{\mathrm{obs}}$ but its own parameters $\tilde{W}_{\mathrm{t}}^{(k)}, \tilde{W}_{\mathrm{s}}^{(k)}$ and $\tilde{b}$.

The two views thus treat the level of the window differently. The observation view sends it through the trend maps, where the network is free to re-express how the level evolves, whereas the standardized view keeps it out of them and restores it unchanged. Rather than committing to one treatment, we compute both forecasts in full and weigh them step by step (Section 3.5).

3.4 Token-Wise Complementary Experts

Dual-Correlation Attention Each view H^{v}, $v \in \{\mathrm{obs}, \mathrm{std}\}$, is turned into a forecast $\hat{Y}^{v}$ of the endogenous variables by its own stack of encoder layers, whose dual-correlation attention (Qiu et al., 2026) models temporal and channel correlations in two branches. In every encoder layer of both branches and both views, this attention lets each token absorb its correlated context, and the Token-Wise Complementary Experts (TCE) then transform the token.

Expert Families Although attention has supplied each token with its correlated context, tokens still differ in content: periodic and non-periodic structures call for different inductive biases (Dong et al., 2025), whereas a transformation shared by all tokens applies the same processing logic to each of them (Ortigossa et al., 2026). We therefore let the content of each token decide the kind of function applied to it. As shown in Figure 2b, a token $x \in \mathbb{R}^{d}$ entering TCE, with d the model dimension, is offered two families of experts that differ in functional form rather than in weights alone. Every expert maps x through a hidden width d_{ff} back to $\mathbb{R}^{d}$, without bias terms and with dropout before its last layer. A frequency expert stacks two Fourier-analysis layers (Dong et al., 2025), which carry a periodic inductive bias, whereas a time expert is a standard two-layer perceptron, which retains general-purpose nonlinear mapping without a periodic prior:

$$\mathcal{E}^{\mathrm{freq}}(x) = \phi_2\big(\phi_1(x)\big), \quad \phi_l(u) = \big[\cos(W^{l}u) \,\|\, \sin(W^{l}u) \,\|\, \mathrm{GELU}(\overline{W}^{l}u)\big], \tag{5}$$

$$\mathcal{E}^{\mathrm{time}}(x) = W^{\mathrm{out}}\,\mathrm{GELU}\big(W^{\mathrm{in}}x\big), \tag{6}$$

where $\|$ denotes concatenation, W^{l} and $\overline{W}^{l}$ project to one quarter and one half of the output width of ϕ_l, which is d_{ff} for $l = 1$ and d for $l = 2$, and $W^{\mathrm{in}} \in \mathbb{R}^{d_{\mathrm{ff}}\times d}$ and $W^{\mathrm{out}} \in \mathbb{R}^{d\times d_{\mathrm{ff}}}$.

Routing and Gating The E routed experts are divided equally, $E/2$ per family, and each family is routed independently with top-k selection (Shazeer et al., 2017),

$$p^{f} = \mathrm{Softmax}\big(W_{\mathrm{r}}^{f}x/\tau\big), \quad o^{f} = \sum_{i\in\mathcal{A}^{f}} \frac{p_i^{f}}{\sum_{j\in\mathcal{A}^{f}} p_j^{f}}\,\mathcal{E}_i^{f}(x), \tag{7}$$

where $f \in \{\text{freq}, \text{time}\}$ indexes the two families, $\mathcal{E}_i^f$ is the i-th expert within family f, $W_{\text{r}}^f \in \mathbb{R}^{(E/2)\times d}$ is the router matrix of family f, τ is the router temperature, and $\mathcal{A}^f = \text{TopK}(p^f)$ is the set of experts activated for the token. A gate reads the same token and mixes the two family outputs, and an always-on shared expert captures knowledge common to all tokens (Dai et al., 2024; Ortigossa et al., 2026). Unlike the routed experts, the shared expert acts on the whole sequence $Z \in \mathbb{R}^{d\times P}$ of the P tokens in the encoder layer: a depthwise convolution with kernel size 3 first mixes each token with its two neighbors in every feature channel, and two pointwise projections of the same shapes as in a time expert follow, $E_{\text{sh}} = W_{\text{sh}}^{\text{out}}\,\text{GELU}\big(W_{\text{sh}}^{\text{in}}\,\text{DWConv}_3(Z)\big)$. With e_{sh} the column of E_{sh} at the position of x,

$$g = \text{Sigmoid}\big(w_g^\top x\big)\,, \quad o = g\,o^{\text{freq}} + (1-g)\,o^{\text{time}}, \tag{8}$$

$$\text{TCE}(x) = o + \text{Sigmoid}\big(w_s^\top x\big)\,e_{\text{sh}}, \tag{9}$$

where $w_g, w_s \in \mathbb{R}^d$ are the family and shared-expert gate vectors, and w_g is initialized at zero so that both families start with equal weight. A periodic token can thus lean on the frequency family, a non-periodic one on the time family, and a mixed one on both (Figure 1b).

3.5 Horizon-Wise View Reconciliation

Sections 3.3–3.4 produce two forecasts of the endogenous variables, $\hat{Y}^{\text{obs}}$ and $\hat{Y}^{\text{std}} \in \mathbb{R}^{N\times F}$, each supervised in its own right (Section 3.6). They are reconciled position by position along the forecast horizon,

$$\hat{Y}^{\text{endo}}_{:,h} = \omega_h\,\hat{Y}^{\text{obs}}_{:,h} + (1-\omega_h)\,\hat{Y}^{\text{std}}_{:,h}, \quad h = 1,\dots,F, \tag{10}$$

where $\omega = \text{Sigmoid}(\psi)$ and $\psi \in \mathbb{R}^F$ is learned, initialized at zero so that the two views start with equal weight at every step. Unlike the family gate g, ω does not read the input, so ψ_h accumulates a preference belonging to the step rather than to any one window.

3.6 Loss Function

All supervised terms use the same discrepancy $\ell(\cdot,\cdot)$, taken to be the mean absolute error. Besides the forecasting loss on the reconciled output, every view is supervised as a complete predictor, and every branch is supervised on the auxiliary target through which it discovers its correlation:

$$\mathcal{L}_{\text{pred}} = \ell\Big(\hat{Y}^{\text{endo}}, Y^{\text{endo}}\Big)\,, \quad \mathcal{L}_{\text{view}} = \sum_v \ell\Big(\hat{Y}^{v}, Y^{\text{endo}}\Big)\,, \tag{11}$$

$$\mathcal{L}_{\text{cor}} = \sum_v \Big[\ell\Big(\hat{Y}^{\text{exo},v}, Y^{\text{exo}}\Big) + \ell\Big(\hat{H}^{v}_{\text{endo}}, H^{v}_{\text{endo}}\Big)\Big]\,, \tag{12}$$

where v ranges over the two views, $\hat{Y}^{\text{exo},v}$ is the temporal branch's forecast of the future exogenous series, and $\hat{H}^{v}_{\text{endo}}$ is the channel branch's reconstruction of H^{v}_{endo}, the endogenous part of H^v, from its exogenous part. Routing is regularized by a load-balancing term (Fedus et al., 2022) applied only inside each family and weighted by the share of the gate the family receives, so that one family can be turned down wherever its form adds nothing,

$$\mathcal{L}_{\text{bal}} = \underset{f,\,l}{\text{mean}}\ \ \bar{g}_f\,\frac{E}{2}\sum_{i=1}^{E/2} c_i\,\bar{p}_i, \tag{13}$$

averaged over the families f of every encoder layer l. Here c_i is the share of the family's assignments received by its i-th expert, that is, $c_i = \sum_{r=1}^{R}\mathbf{1}[i\in\mathcal{A}_r^f]\big/\sum_{r=1}^{R}|\mathcal{A}_r^f|$ over the R tokens of a batch, $\bar{p}_i$ is its mean router weight, and $\bar{g}_f$ the average share the family gate assigns to f. The model is trained end to end on

$$\mathcal{L} = \mathcal{L}_{\text{pred}} + \gamma\,\mathcal{L}_{\text{view}} + \beta\,\mathcal{L}_{\text{cor}} + \lambda\,\mathcal{L}_{\text{bal}}, \tag{14}$$

where γ, β and λ are weights that balance the contribution of the two view predictors, of correlation modeling, and of routing balance against the forecasting loss.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETTINGS

Datasets We evaluate CVENet on twelve public real-world datasets that satisfy the conditions of forecasting with exogenous variables, comprising five electricity-price datasets from the EPF benchmark (Lago et al., 2021; Wang et al., 2024) and seven datasets released with DAG (Qiu et al., 2026). In all of them the future exogenous variables are either known in advance or can be forecast with high accuracy, and each is evaluated under both regimes, with and without those values at inference time. We additionally report results on EAST-POINT127, a tokamak dataset in which forty actuator signals drive eleven electron-density channels. Further details of the datasets are provided in Appendix A.1.

Baselines We compare CVENet against ten baselines: 1) methods that inherently support future exogenous variables, namely DAG (Qiu et al., 2026), GCGNet (Li et al., 2026b), TimeXer (Wang et al., 2024), TFT (Lim et al., 2021) and TiDE (Das et al., 2023); and 2) methods that do not, namely DUET (Qiu et al., 2025b), CrossLinear (Zhou et al., 2025), Amplifier (Fei et al., 2025), TimeKAN (Huang et al., 2025) and PatchTST (Nie et al., 2023). Following DAG (Qiu et al., 2026), the second group is extended with an MLP fusion approach when future exogenous variables are available, so that every model receives the same inputs (Algorithm 1 in Appendix A.4). On EAST-POINT127 we compare against DAG, TimeXer, TFT and TiDE.

Implementation Details To keep consistent with previous works, we adopt Mean Squared Error (MSE) and Mean Absolute Error (MAE) as evaluation metrics; on EAST-POINT127 we additionally report the Normalized Root Mean Squared Error (NRMSE) and the coefficient of determination R^2 (Appendix A.2). We conduct both short-term and long-term prediction experiments. For Colbun and Rapel, the short-term setting uses a lookback window of 60 with a prediction horizon of 10, and the long-term setting a lookback window of 180 with a horizon of 30; for the other ten public datasets, the short-term setting uses a lookback window of 168 with a horizon of 24, and the long-term setting a lookback window of 720 with a horizon of 360. On EAST-POINT127 the lookback window is fixed to 96 and the horizons are 96, 192, 336 and 720. Further implementation details are provided in Appendix A.

Table 1: Average results on 12 real-world datasets that satisfy the TSF-X task, where the inputs are (X^{endo}, X^{exo}, and Y^{exo}). **Red**: the best, Blue: the 2nd best. Full results are available in Table 7 of Appendix B.2.

Models	CVENet		DAG		GCGNet		TimeXer		TFT		TiDE		DUET		CrossLinear		Amplifier		TimeKAN		PatchTST	
Metrics	mse	mae	mse	mae	mse	mae	mse	mae	mse	mae	mse	mae	mse	mae	mse	mae	mse	mae	mse	mae	mse	mae
NP	**0.342**	**0.330**	0.362	0.344	0.370	0.348	0.418	0.371	0.379	0.375	0.443	0.400	0.411	0.408	0.371	0.387	0.420	0.418	0.405	0.419	0.390	0.396
PJM	**0.081**	**0.176**	0.093	0.180	0.095	0.187	0.108	0.198	0.114	0.207	0.142	0.246	0.102	0.197	0.112	0.223	0.137	0.246	0.139	0.262	0.133	0.259
BE	0.438	0.289	**0.423**	**0.279**	0.431	0.294	0.452	0.290	0.454	0.291	0.498	0.325	0.515	0.354	0.479	0.337	0.559	0.413	0.548	0.407	0.577	0.432
FR	**0.410**	**0.218**	0.414	0.219	0.415	0.234	0.427	0.241	0.504	0.257	0.484	0.281	0.496	0.327	0.483	0.298	0.554	0.408	0.547	0.374	0.588	0.410
DE	0.372	**0.365**	**0.370**	0.370	0.401	0.389	0.475	0.418	0.489	0.446	0.499	0.447	0.482	0.430	0.485	0.452	0.473	0.441	0.473	0.445	0.501	0.455
Energy	0.126	0.268	**0.124**	**0.267**	0.131	0.277	0.163	0.315	0.130	0.283	0.153	0.302	0.203	0.367	0.239	0.402	0.233	0.389	0.218	0.381	0.226	0.377
Sdwpfm1	**0.390**	**0.441**	0.423	0.461	0.424	0.457	0.701	0.609	0.482	0.474	0.483	0.507	0.599	0.570	0.426	0.502	0.437	0.490	0.447	0.534	0.435	0.502
Sdwpfm2	**0.430**	**0.466**	0.477	0.485	0.475	0.486	0.803	0.653	0.476	0.488	0.486	0.516	0.514	0.490	0.533	0.573	0.491	0.512	0.497	0.564	0.510	0.547
Sdwpfh1	**0.436**	**0.486**	0.448	0.486	0.450	0.500	0.746	0.643	0.479	0.491	0.453	0.508	0.539	0.516	0.557	0.593	0.537	0.598	0.577	0.638	0.468	0.527
Sdwpfh2	**0.503**	**0.520**	0.523	0.530	0.520	0.536	0.891	0.719	0.566	0.521	0.599	0.583	0.647	0.566	0.538	0.574	0.521	0.581	0.647	0.672	0.549	0.580
Colbun	**0.089**	**0.128**	0.098	0.154	0.107	0.175	0.145	0.235	0.238	0.297	0.164	0.227	0.198	0.266	0.126	0.195	0.173	0.246	0.128	0.175	0.239	0.309
Rapel	**0.212**	**0.272**	0.230	0.305	0.306	0.307	0.344	0.362	0.305	0.333	0.320	0.351	0.269	0.326	0.252	0.313	0.257	0.321	0.249	0.311	0.269	0.332
1st Count	**9**	**10**	3	2	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

4.2 MAIN RESULTS

Comprehensive forecasting results are presented in Table 1 and Table 2 to demonstrate the performance of CVENet on forecasting with exogenous variables. We have the following observations:

1) Compared with forecasters of various architectures, CVENet achieves outstanding predictive performance on the twelve public datasets. As shown in Table 1, CVENet ranks first on 9 of the 12 datasets in MSE and on 10 of the 12 in MAE, and in Table 7 it obtains 23 first-place rankings in MSE and 26 in MAE, demonstrating the advantages of retaining complementary views of the history and adapting the functional form to each token. Averaged over the twelve datasets, CVENet reduces the

Table 2: Results on the EAST-POINT127 dataset. **Red**: the best, Blue: the 2nd best. Avg means the average results over the four forecasting horizons. For R^2 a higher value is better; for all other metrics a lower value is better.

Models		CVENet				DAG				TimeXer				TFT				TiDE			
Metrics		mse	mae	nrmse	R^2	mse	mae	nrmse	R^2	mse	mae	nrmse	R^2	mse	mae	nrmse	R^2	mse	mae	nrmse	R^2
EAST-POINT127	96	**0.024**	0.078	**0.099**	**0.989**	0.026	0.078	0.099	0.988	0.025	0.079	0.099	0.989	0.044	0.101	0.126	0.980	0.027	**0.077**	0.099	0.988
	192	**0.055**	**0.124**	**0.154**	**0.975**	0.067	0.134	0.167	0.969	0.066	0.131	0.163	0.970	0.120	0.191	0.232	0.945	0.101	0.178	0.218	0.954
	336	**0.112**	**0.187**	**0.229**	**0.949**	0.138	0.210	0.257	0.937	0.140	0.207	0.255	0.936	0.233	0.279	0.357	0.894	0.199	0.243	0.294	0.909
	720	**0.289**	**0.322**	**0.387**	**0.868**	0.331	0.357	0.429	0.849	0.432	0.366	0.454	0.803	0.796	0.564	0.671	0.637	0.761	0.530	0.645	0.653
	Avg	**0.120**	**0.178**	**0.217**	**0.945**	0.140	0.195	0.238	0.936	0.166	0.196	0.243	0.924	0.298	0.284	0.346	0.864	0.272	0.257	0.314	0.876
1st Count		**5**	**4**	**5**	**5**	0	0	0	0	0	0	0	0	0	0	0	0	0	1	0	0

MSE and MAE of DAG, the strongest baseline, by 3.9% and 3.0%, with the largest reductions of 12.9% in MSE on PJM and 16.9% in MAE on Colbun.

2) On EAST-POINT127 (Table 2), CVENet ranks first in 19 of the 20 entries, the only exception being MAE at the shortest horizon, and its advantage becomes more pronounced at longer forecasting horizons. Averaged over the four horizons, it reduces the MSE of DAG by 14.3% and raises R^2 from 0.936 to 0.945.

Table 3: Ablation studies for CVENet. The best results are highlighted in **bold**.

Dataset	NP		DE		Energy		EAST-POINT127		Average	
Metrics	mse	mae	mse	mae	mse	mae	mse	mae	mse	mae
(a) w/o Token-Wise Complementary Experts	0.631	0.544	0.516	0.436	0.216	0.367	0.308	0.328	0.418	0.419
(b) w/o Complementary View	0.585	0.492	0.573	0.475	0.268	0.407	0.337	0.339	0.441	0.428
(c) w/o Horizon-Wise View Reconciliation	0.520	0.449	0.494	0.436	0.254	0.401	0.296	0.325	0.391	0.403
CVENet	**0.503**	**0.436**	**0.473**	**0.421**	**0.183**	**0.334**	**0.289**	**0.322**	**0.362**	**0.378**

4.3 Model Analyses

Ablation Studies To ascertain the impact of different components within CVENet, we compare the full model with the following variants: (a) Replacing the Token-Wise Complementary Experts with a standard feed-forward network, which also removes the load-balancing term. (b) Removing the standardized view, so that the forecast of the observation view serves directly as the output. (c) Replacing the horizon-wise weights ω with a single learned weight shared by all forecast steps. EAST-POINT127 uses a lookback length of 96 and a horizon of 720, while NP, DE and Energy use a lookback length of 720 and a horizon of 360; the results are shown in Table 3. We make the following observations:

• Replacing the Token-Wise Complementary Experts degrades performance on every dataset, most markedly on NP, highlighting the benefit of complementary experts.

• Removing the standardized view causes the largest degradation on average, with the largest relative increases in both MSE and MAE on Energy, showing that committing to a single treatment of non-stationarity is costly.

• Sharing one reconciliation weight across the horizon also performs worse on every dataset, with the widest gap on Energy, although it still outperforms the single-view variant throughout. Since variant (c) differs from the full model only in the $F - 1$ additional scalar parameters that let the weight vary with the forecast step, this consistent gap supports step-dependent reconciliation.

• Ultimately, the full CVENet, which combines all three components, achieves the best results on all four datasets, including the tokamak dataset EAST-POINT127, demonstrating that each component contributes to the forecasting performance and that their gains are complementary. Appendix B.3 and Appendix B.4 further show that switching off the view loss gives the largest error on every dataset and that CVENet attains a lower MAE than the three compared baselines at every look-back window.

Forecasting without Future Exogenous Variables Future exogenous variables are not available in every deployment, so we repeat the evaluation with only the historical endogenous and exogenous series as input, in which case the channel branch of CVENet consumes the future exogenous series forecast by its temporal branch; the average results are reported in Table 4 and the full results in Table 8. CVENet again ranks first on 7 of the 12 datasets in both MSE and MAE, with 18 first-place rankings in MSE and 22 in MAE in Table 8.

Table 4: Average results under the setting where only historical exogenous variables are used, without future exogenous variables. The inputs are (X^{endo} and X^{exo}). **Red**: the best, Blue: the 2nd best. Full results are available in Table 8 of Appendix B.2.

Models	CVENet		DAG		GCGNet		TimeXer		TFT		TiDE		DUET		CrossLinear		Amplifier		TimeKAN		PatchTST	
Metrics	mse	mae	mse	mae	mse	mae	mse	mae	mse	mae	mse	mae	mse	mae	mse	mae	mse	mae	mse	mae	mse	mae
NP	0.432	0.383	**0.419**	**0.380**	0.430	0.381	0.440	0.383	0.647	0.488	0.467	0.416	0.444	0.383	0.451	0.394	0.520	0.448	0.484	0.435	0.457	0.401
PJM	**0.121**	**0.204**	0.126	0.210	0.131	0.216	0.141	0.229	0.200	0.270	0.158	0.253	0.140	0.226	0.147	0.241	0.152	0.245	0.175	0.270	0.148	0.243
BE	0.462	0.307	0.462	**0.297**	**0.459**	0.306	0.477	0.301	0.563	0.354	0.547	0.348	0.473	0.305	0.477	0.300	0.502	0.325	0.488	0.315	0.485	0.309
FR	0.448	0.258	**0.435**	0.249	0.435	**0.246**	0.454	0.247	0.535	0.288	0.494	0.290	0.468	0.262	0.476	0.257	0.494	0.286	0.491	0.276	0.470	0.264
DE	**0.592**	**0.478**	0.603	0.487	0.614	0.501	0.659	0.507	0.684	0.515	0.644	0.519	0.660	0.513	0.635	0.508	0.712	0.548	0.669	0.526	0.696	0.527
Energy	0.154	0.302	0.150	0.300	0.154	0.305	0.172	0.326	0.376	0.480	0.162	0.311	0.160	0.308	0.201	0.336	0.180	0.327	**0.147**	**0.296**	0.203	0.341
Sdwpfm1	**0.676**	**0.578**	0.689	0.604	0.711	0.595	0.701	0.609	0.984	0.728	0.713	0.633	0.724	0.583	0.809	0.694	0.843	0.685	0.725	0.672	0.733	0.651
Sdwpfm2	**0.773**	**0.620**	0.786	0.657	0.807	0.643	0.803	0.653	1.044	0.767	0.816	0.682	0.820	0.641	0.800	0.687	0.942	0.732	0.836	0.701	0.834	0.714
Sdwpfh1	**0.731**	0.629	0.733	0.643	0.741	**0.628**	0.746	0.643	0.793	0.698	0.808	0.699	0.779	0.644	0.768	0.702	1.100	0.820	0.804	0.720	0.804	0.717
Sdwpfh2	**0.849**	**0.684**	0.870	0.695	0.886	0.701	0.891	0.719	0.926	0.746	0.919	0.751	1.007	0.715	0.956	0.774	0.980	0.768	0.941	0.782	1.025	0.802
Colbun	0.122	**0.155**	**0.117**	0.155	0.119	0.165	0.132	0.219	0.556	0.386	0.188	0.240	0.147	0.198	0.129	0.195	0.152	0.210	0.201	0.253	0.150	0.221
Rapel	**0.238**	**0.277**	0.261	0.281	0.265	0.284	0.344	0.362	0.308	0.334	0.323	0.357	0.304	0.306	0.240	0.299	0.337	0.348	0.342	0.337	0.325	0.334
1st Count	**7**	**7**	3	2	1	2	0	0	0	0	0	0	0	0	0	0	0	0	1	1	0	0

4.4 Visualization of Predictions

Figure 3 visualizes the predictions of CVENet, DAG and GCGNet on the NP dataset, where CVENet follows the ground truth most closely both without and with future exogenous variables (Figure 3d–f and g–i). More visualization results are provided in Appendix B.1.

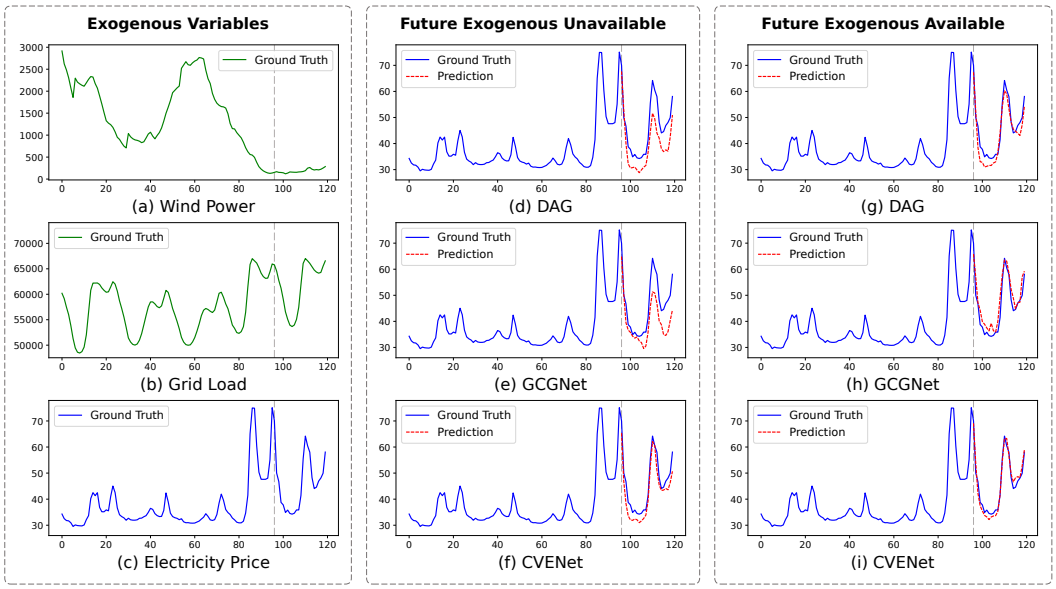


Figure 3: Prediction results on the NP dataset. (a–b) Exogenous variables: wind power and grid load. (c) Endogenous variables: Electricity Price. (d–f) DAG, GCGNet and CVENet predictions when future exogenous variables are unavailable. (g–i) DAG, GCGNet and CVENet predictions when future exogenous variables are available.

5 Conclusion

In this paper, we propose **CVENet**, a complementary view and expert network for time series forecasting with exogenous variables. Complementary View Construction keeps two treatments of non-stationarity as separate views; Token-Wise Complementary Experts mix a frequency-domain and a time-domain expert family for each token; and Horizon-Wise View Reconciliation combines the two view forecasts step by step. Rather than fixing the treatment of non-stationarity and the functional form applied to each token in advance, CVENet thus resolves the former along the forecast horizon and the latter within each token. Extensive experiments on twelve real-world datasets, with and without future exogenous variables, and on a tokamak dataset demonstrate that CVENet achieves state-of-the-art forecasting performance.

AI USE STATEMENT

We used generative AI tools throughout this work: to refine the idea and the experimental design, to implement the method, to process and interpret results, to search related literature, to produce figures and format references, and to draft and polish sections of this paper. Tasks involving synthetic data generation and mathematical proofs are not applicable to this work. We have reviewed all AI-assisted work and take responsibility for the final content of this paper, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

Our work addresses the methodological problem of time series forecasting and poses no foreseeable ethical risk. All datasets used are machine-recorded measurements that involve no human subjects and contain no personally identifiable information.

REPRODUCIBILITY STATEMENT

The model architecture is formalized with equations in the main text, and further implementation details are provided in Appendix A. The twelve public benchmarks are available from their original sources, whereas the EAST-POINT127 recordings cannot be publicly released. The source code is provided in the supplementary material and will be made public once the paper is accepted.

REFERENCES

Francisco Martinez Alvarez, Alicia Troncoso, Jose C Riquelme, and Jesus S Aguilar Ruiz. Energy time series forecasting based on pattern sequence similarity. *IEEE Transactions on Knowledge and Data Engineering*, 23(8):1230–1243, 2011.

Damai Dai, Chengqi Deng, Chenggang Zhao, RX Xu, Huazuo Gao, Deli Chen, Jiashi Li, Wangding Zeng, Xingkai Yu, Y Wu, et al. DeepSeekMoE: Towards ultimate expert specialization in mixture-of-experts language models. In *ACL*, 2024.

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.

Yihong Dong, Ge Li, Yongding Tao, Xue Jiang, Kechi Zhang, Jia Li, Jinliang Deng, Jing Su, Jun Zhang, and Jingjing Xu. FAN: Fourier analysis networks. In *NeurIPS*, 2025.

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*, volume 37, pp. 7522–7529, 2023.

William Fedus, Barret Zoph, and Noam Shazeer. Switch Transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(1):5232–5270, 2022.

Jingru Fei, Kun Yi, Wei Fan, Qi Zhang, and Zhendong Niu. Amplifier: Bringing attention to neglected low-energy components in time series forecasting. In *AAAI*, volume 39, pp. 11645–11653, 2025.

Hans Hersbach, Bill Bell, Paul Berrisford, Shoji Hirahara, András Horányi, Joaquín Muñoz-Sabater, Julien Nicolas, Carole Peubey, Raluca Radu, Dinand Schepers, et al. The ERA5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146(730):1999–2049, 2020.

Yifan Hu, Peiyuan Liu, Peng Zhu, Dawei Cheng, and Tao Dai. Adaptive multi-scale decomposition framework for time series forecasting. In *AAAI*, volume 39, pp. 17359–17367, 2025.

Songtao Huang, Zhen Zhao, Can Li, and Lei Bai. TimeKAN: KAN-based frequency decomposition learning architecture for long-term time series forecasting. In *ICLR*, 2025.

Xuanwen Huang, Yang Yang, Yang Wang, Chunping Wang, Zhisheng Zhang, Jiarong Xu, Lei Chen, and Michalis Vazirgiannis. DGraph: A large-scale financial dataset for graph anomaly detection. In *NeurIPS*, volume 35, pp. 22765–22777, 2022.

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

Julian Kates-Harbeck, Alexey Svyatkovskiy, and William Tang. Predicting disruptive instabilities in controlled fusion plasmas through deep learning. *Nature*, 568(7753):526–531, 2019.

Taesung Kim, Jinhee Kim, Yunwon Tae, Cheonbok Park, Jang-Ho Choi, and Jaegul Choo. Reversible instance normalization for accurate time-series forecasting against distribution shift. In *ICLR*, 2022.

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.

Hao Li, Lu Zhang, Liu Chong, Yankai Chen, Pengyang Wang, and Yingjie Zhou. SeesawNet: Towards non-stationary time series forecasting with balanced modeling of common and specific dependencies. In *IJCAI*, 2026a.

Zhe Li, Xiangfei Qiu, Peng Chen, Yihang Wang, Hanyin Cheng, Yang Shu, Jilin Hu, Chenjuan Guo, Aoying Zhou, Qingsong Wen, et al. TSFM-Bench: A comprehensive and unified benchmark of foundation models for time series forecasting. In *SIGKDD*, 2025.

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*, 2026b.

H. Lian, H.Q. Liu, K. Li, Z.Y. Zou, J.P. Qian, M.Q. Wu, G.Q. Li, L. Zeng, Q. Zang, B. Lv, Y.X. Jie, and EAST Team. EAST kinetic equilibrium reconstruction combining with polarimeter-interferometer internal measurement constraints. *Journal of Instrumentation*, 12(12):C12036, 2017.

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.

Jiafeng Lin, Mengren Zheng, Simeng Ye, Yuxuan Wang, Huan Zhang, Yuhui Liu, Zhongyi Pei, and Jianmin Wang. Aura: Universal multi-dimensional exogenous integration for aviation time series. In *SIGKDD*, pp. 7645–7655, 2026.

Shengsheng Lin, Weiwei Lin, Xinyi Hu, Wentai Wu, Ruichao Mo, and Haocheng Zhong. CycleNet: Enhancing time series forecasting through modeling periodic patterns. In *NeurIPS*, volume 37, pp. 106315–106345, 2024.

Yunfei Ling, Jun Du, Zijie Liu, Yao Huang, Yuehang Wang, Bingjia Xiao, and Xin Fang. PaMMA-Net: Plasmas magnetic measurement evolution based on data-driven incremental accumulative prediction. *Nuclear Fusion*, 65(10):106027, 2025.

Peiyuan Liu, Beiliang Wu, Yifan Hu, Naiqi Li, Tao Dai, Jigang Bao, and Shu-Tao Xia. TimeBridge: Non-stationarity matters for long-term time series forecasting. In *ICML*, 2025a.

Yiwen Liu, Chenyu Zhang, Junjie Song, Siqi Chen, Yin Sun, Zihan Wang, Lingming Zeng, Yuji Cao, and Junming Jiao. MoFE-Time: Mixture of frequency domain experts for time-series forecasting models. *arXiv preprint arXiv:2507.06502*, 2025b.

Yong Liu, Haixu Wu, Jianmin Wang, and Mingsheng Long. Non-stationary transformers: Exploring the stationarity in time series forecasting. In *NeurIPS*, volume 35, pp. 9881–9893, 2022.

Zhiding Liu, Mingyue Cheng, Zhi Li, Zhenya Huang, Qi Liu, Yanhu Xie, and Enhong Chen. Adaptive normalization for non-stationary time series forecasting: A temporal slice perspective. In *NeurIPS*, volume 36, pp. 14273–14292, 2023.

Yisheng Lv, Yanjie Duan, Wenwen Kang, Zhengxi Li, and Fei-Yue Wang. Traffic flow prediction with big data: A deep learning approach. *IEEE Transactions on Intelligent Transportation Systems*, 16(2):865–873, 2014.

Yuqi Nie, Nam H Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. A time series is worth 64 words: Long-term forecasting with transformers. In *ICLR*, 2023.

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.

Evandro S Ortigossa, Guy Lutsker, and Eran Segal. MoHETS: Long-term time series forecasting with mixture-of-heterogeneous-experts. *arXiv preprint arXiv:2601.21866*, 2026.

Zhicheng Pan, Yihang Wang, Yingying Zhang, Sean Bin Yang, Yunyao Cheng, Peng Chen, Chenjuan Guo, Qingsong Wen, Xiduo Tian, Yunliang Dou, et al. MagicScaler: Uncertainty-aware, predictive autoscaling. *Proc. VLDB Endow.*, 16(12):3808–3821, 2023.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. PyTorch: An imperative style, high-performance deep learning library. In *NeurIPS*, 2019.

Xiangfei Qiu, Jilin Hu, Lekui Zhou, Xingjian Wu, Junyang Du, Buang Zhang, Chenjuan Guo, Aoying Zhou, Christian S. Jensen, Zhenli Sheng, and Bin Yang. TFB: towards comprehensive and fair benchmarking of time series forecasting methods. *Proc. VLDB Endow.*, 17(9):2363–2377, 2024.

Xiangfei Qiu, Zhe Li, Wanghui Qiu, Shiyan Hu, Lekui Zhou, Xingjian Wu, Zhengyu Li, Chenjuan Guo, Aoying Zhou, Zhenli Sheng, Jilin Hu, Christian S. Jensen, and Bin Yang. TAB: Unified benchmarking of time series anomaly detection methods. In *Proc. VLDB Endow.*, pp. 2775–2789, 2025a.

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, 2025b.

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 *ICML*, 2026.

Cécile Rousseau, Samuel Jackson, Rodrigo H. Ordonez-Hurtado, Nicola C. Amorisco, Tobia Boschi, George K. Holt, Andrea Loreti, Eszter Székely, Alexander Whittle, Adriano Agnello, Stanislas Pamela, Alessandra Pascale, Robert Akers, Juan Bernabe Moreno, Sue Thorne, and Mykhaylo Zayats. TokaMark: A comprehensive benchmark for MAST tokamak plasma models. *arXiv preprint arXiv:2602.10132*, 2026.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc V Le, Geoffrey E Hinton, and Jeff Dean. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. In *ICLR*, 2017.

Xiaoming Shi, Shiyu Wang, Yuqi Nie, Dianqi Li, Zhou Ye, Qingsong Wen, and Ming Jin. Time-MoE: Billion-scale time series foundation models with mixture of experts. In *ICLR*, 2025.

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

Yanru Sun, Zongxia Xie, Emadeldeen Eldele, Dongyue Chen, Qinghua Hu, and Min Wu. Learning pattern-specific experts for time series forecasting under patch-level distribution shift. In *NeurIPS*, 2025.

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.

Chenguang Wan, Zhi Yu, Feng Wang, Xiaojuan Liu, and Jiangang Li. Experiment data-driven modeling of tokamak discharge in EAST. *Nuclear Fusion*, 61(6):066015, 2021.

Chenguang Wan, Feda Almuhisen, Philippe Moreau, Rémy Nouailletas, Zhisong Qu, Youngwoo Cho, Robin Varennes, Kyungtak Lim, Kunpeng Li, Jia Huang, Weidong Chen, Jiangang Li, Xavier Garbet, and WEST Team. Machine learning prediction of plasma behavior from discharge configurations on WEST. *Nuclear Fusion*, 66(4):044001, 2026.

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, 2024.

Billy M Williams. Multivariate vehicular traffic flow prediction: Evaluation of ARIMAX modeling. *Transportation Research Record*, 1776(1):194–200, 2001.

Haixu Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. In *NeurIPS*, volume 34, pp. 22419–22430, 2021.

Guoqi Yu, Jing Zou, Xiaowei Hu, Angelica I Aviles-Rivero, Jing Qin, and Shujun Wang. Revitalizing multivariate time series forecasting: Learnable decomposition with inter-series dependencies and intra-series variations modeling. In *ICML*, 2024.

Ailing Zeng, Muxi Chen, Lei Zhang, and Qiang Xu. Are transformers effective for time series forecasting? In *AAAI*, volume 37, pp. 11121–11128, 2023.

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*, pp. 4120–4131, 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*, volume 162, pp. 27268–27286, 2022.

Jiawen Zhu, Shuhan Liu, Di Weng, and Yingcai Wu. Dynamic TMoE: A drift-aware dynamic mixture of experts framework for non-stationary time series forecasting. In *ICML*, 2026.

A Experimental Details

A.1 Datasets

Table 5: Statistics of datasets. Ex. and En. are abbreviations for the exogenous variables and endogenous variables, respectively.

Dataset	#Num	Sampling Frequency	Lengths	Split	Ex. Descriptions	En. Descriptions
NP	2	1 Hour	52,416	7:1:2	Grid Load, Wind Power	Nord Pool Electricity Price
PJM	2	1 Hour	52,416	7:1:2	System Load, SyZonal COMED load	Pennsylvania-New Jersey-MarylandElectricity Price
BE	2	1 Hour	52,416	7:1:2	Generation, System Load	Belgium's Electricity Price
FR	2	1 Hour	52,416	7:1:2	Generation, System Load	France's Electricity Price
DE	2	1 Hour	52,416	7:1:2	Wind power, Ampirion zonal load	German's Electricity Price
Energy	5	1 Hour	13,064	7:1:2	Battery, Geothermal, Hydroelectric, Solar, Wind	Thermoelectric
Colbún	2	1 Day	2,958	7:1:2	Precipitation, Tributary Inflow	Water Level
Rapel	2	1 Day	3,366	7:1:2	Precipitation, Tributary Inflow	Water Level
Sdwpfh	6	1 Hour	14,641	7:1:2	Climate Feature	Active Power
Sdwpfm	6	30 Minutes	29,281	7:1:2	Climate Feature	Active Power
EAST-POINT127	40	1 Millisecond	1,915,000	7:1:2	Actuator Signal	Electron Density (POINT)

We conduct experiments on 12 real-world datasets with exogenous variables, where future exogenous variables are either approximately known or can be obtained with high accuracy. These include 5 EPF (Lago et al., 2021; Wang et al., 2024) datasets (NP, PJM, BE, FR, DE) and 7 datasets collected by DAG (Qiu et al., 2026) (Energy, Colbún, Rapel, Sdwpfh1, Sdwpfh2, Sdwpfm1, Sdwpfm2). **NP** records Nord Pool market with hourly electricity price as the endogenous variable and grid load and wind power forecast as exogenous features. **PJM** records the Pennsylvania–New Jersey–Maryland Interconnection market with zonal electricity price in the COMED area as the endogenous variable, and system load and COMED load forecasts as exogenous variables. **BE** records Belgium's electricity market with hourly price as the endogenous variable and load forecast in Belgium plus generation forecast from France as exogenous variables. **FR** records the French electricity market with price as the endogenous variable and generation and load forecasts as exogenous variables. **DE** records the German electricity market with hourly price as the endogenous variable and zonal load forecast in Amprion, along with wind and solar generation forecasts as exogenous variables. **Energy** provides hourly power generation data from Chile, with thermoelectric generation as the endogenous variable and battery storage, wind, hydro, and solar as exogenous variables. **Colbún and Rapel** are daily-resolution hydropower datasets with water level as the endogenous variable and precipitation and tributary inflow as exogenous variables. **Sdwpfh1, Sdwpfh2, Sdwpfm1, Sdwpfm2** are wind power datasets from Longyuan wind farm, with active power output (Patv) as the endogenous variable and external weather data from ERA5 (Hersbach et al., 2020) as exogenous variables, including temperature, surface pressure, relative humidity, wind speed, wind direction, and total precipitation.

Table 6: Variables of the EAST-POINT127 dataset.

Type	System	#Channels
Exogenous	Electron cyclotron resonance heating (ECRH)	4
	Ion cyclotron range of frequencies heating (ICRF)	16
	Lower hybrid wave heating and current drive (LHW)	4
	Neutral beam injection (NBI)	8
	Magnetic coil currents	4
	Particle injection	1
	Supersonic molecular beam injection (SMBI)	3
Endogenous	Electron density (POINT)	11

EAST-POINT127 is collected from the Experimental Advanced Superconducting Tokamak (EAST). A tokamak operates in discharges, or shots, during which auxiliary heating and current drive, magnetic coils and fueling systems act on the plasma as actuators. The electron density is regulated mainly through fueling, yet its response is neither instantaneous nor one-to-one, as it depends nonlinearly on wall conditions and transport (Wan et al., 2026). The task on EAST-POINT127 is therefore to forecast the electron density from its own history together with the actuator signals, which are treated as future exogenous variables available over the forecast horizon. The variables are listed in Table 6, and Figure 4 shows representative actuator signals together with the electron density in one discharge. The forty exogenous variables are the actuator signals of seven systems, and the eleven endogenous variables are the electron-density channels measured by the polarimeter-interferometer (POINT) (Lian et al., 2017), all of which are forecast jointly. The dataset contains 127 discharges lasting 7 to 41 seconds each, sampled at one millisecond and concatenated in chronological order into a single series of 1,915,000 steps. The series is split at discharge boundaries, with the first 67 discharges for training, the next 16 for validation and the last 44 for testing, which corresponds to a 7:1:2 split in steps. No discharge is shared between subsets, so the model is evaluated on discharges unseen during training.

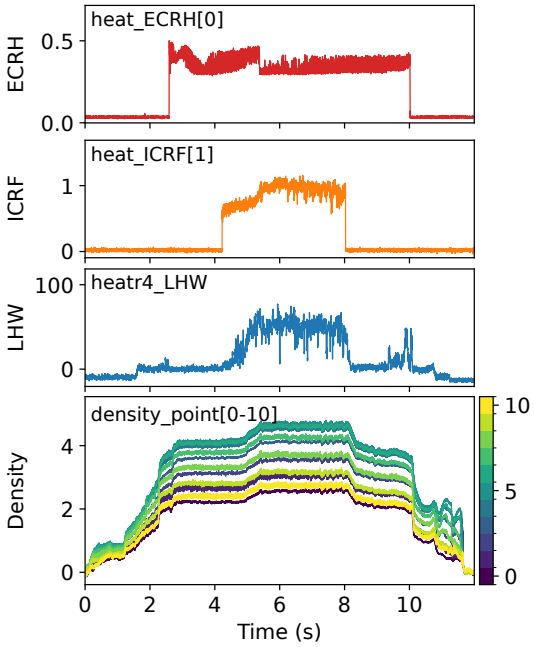


Figure 4: Representative actuator signals and the eleven electron-density channels of one discharge in the EAST-POINT127 dataset.

A.2 Evaluation Metrics

To evaluate model performance, we employ Mean Squared Error (MSE) and Mean Absolute Error (MAE) on all datasets, following the TFB benchmark (Qiu et al., 2024). On EAST-POINT127, we additionally report the Normalized Root Mean Squared Error (NRMSE) and the coefficient of determination (R^2), which are commonly used to evaluate predictions of tokamak plasma behavior (Rousseau et al., 2026; Wan et al., 2026).

Mean Squared Error (MSE) The Mean Squared Error (MSE) measures the average squared difference between the predicted values and the actual observations. As the errors are squared, MSE penalizes large errors more heavily than MAE, which makes it more sensitive to outliers. The mathematical representation of MSE is defined as:

$$\mathrm{MSE} = \frac{1}{N \times F} \sum_{n=1}^{N} \sum_{h=1}^{F} \left(y_h^n - \hat{y}_h^n\right)^2, \tag{15}$$

where y_h^n denotes the ground truth and $\hat{y}_h^n$ denotes the predicted value.

Mean Absolute Error (MAE) The Mean Absolute Error (MAE) measures the average magnitude of the errors without considering their direction, that is, the average absolute difference between the predicted values and the actual observations. The mathematical representation of MAE is given by:

$$\mathrm{MAE} = \frac{1}{N \times F} \sum_{n=1}^{N} \sum_{h=1}^{F} \left|y_h^n - \hat{y}_h^n\right|, \tag{16}$$

where N represents the number of endogenous variables and F represents the number of forecast steps.

Algorithm 1 MLP Fusion Approach

Input: model parameters θ_{Model}, MLP parameters θ_{MLP}; learning rate η; iterations I; training dataset $\mathcal{S}$; historical endogenous variables X^{endo}; historical exogenous variables X^{exo}; future exogenous variables Y^{exo}

Output: prediction $\hat{Y}^{\text{endo}}$

Training Phase:
For $i = 1$ to I do
  For $(X^{\text{endo}}_{\text{train}}, X^{\text{exo}}_{\text{train}}, Y^{\text{endo}}_{\text{train}}, Y^{\text{exo}}_{\text{train}})$ in $\mathcal{S}$
    $z \leftarrow \theta_{\text{Model}}(X^{\text{endo}}_{\text{train}}, X^{\text{exo}}_{\text{train}})$
    $\hat{Y}^{\text{endo}} \leftarrow \theta_{\text{MLP}}(\text{Concat}(z, Y^{\text{exo}}_{\text{train}}))$
    $\mathcal{L} \leftarrow Loss(\hat{Y}^{\text{endo}}, Y^{\text{endo}}_{\text{train}})$
    $\theta_{\text{Model}} \leftarrow \theta_{\text{Model}} - \eta \frac{\partial \mathcal{L}}{\partial \theta_{\text{Model}}},\ \theta_{\text{MLP}} \leftarrow \theta_{\text{MLP}} - \eta \frac{\partial \mathcal{L}}{\partial \theta_{\text{MLP}}}$
  EndFor
EndFor
Inference Phase:
$z \leftarrow \theta_{\text{Model}}(X^{\text{endo}}, X^{\text{exo}})$
$\hat{Y}^{\text{endo}} \leftarrow \theta_{\text{MLP}}(\text{Concat}(z, Y^{\text{exo}}))$
return $\hat{Y}^{\text{endo}}$

Normalized Root Mean Squared Error (NRMSE) The Normalized Root Mean Squared Error (NRMSE) measures the root mean squared error relative to the standard deviation of each signal, which makes the error dimensionless and comparable across signals of different magnitudes, such as the eleven electron-density channels of EAST-POINT127. Since every variable is standardized by the mean and standard deviation of its training split before the errors are computed, the root mean squared error is already normalized, and NRMSE averages it over the M rolling forecast windows:

$$\text{NRMSE} = \frac{1}{M} \sum_{m=1}^{M} \sqrt{\frac{1}{N \times F} \sum_{n=1}^{N} \sum_{h=1}^{F} (y_h^{n,m} - \hat{y}_h^{n,m})^2}, \tag{17}$$

where $y_h^{n,m}$ and $\hat{y}_h^{n,m}$ denote the ground truth and the predicted value in the m-th window.

Coefficient of Determination (R^2) The coefficient of determination (R^2) measures the proportion of the variance in the actual observations that is explained by the predictions. It equals 1 for perfect predictions and 0 for predicting the mean of the observations, and a higher value is better. The mathematical representation of R^2 is defined as:

$$R^2 = 1 - \frac{\sum_{m=1}^{M} \sum_{n=1}^{N} \sum_{h=1}^{F} (y_h^{n,m} - \hat{y}_h^{n,m})^2}{\sum_{m=1}^{M} \sum_{n=1}^{N} \sum_{h=1}^{F} (y_h^{n,m} - \bar{y})^2}, \tag{18}$$

where $\bar{y}$ denotes the mean of the actual observations over all windows, variables and forecast steps.

A.3 Implementation Details

All experiments are conducted with PyTorch (Paszke et al., 2019) in Python 3.8 on a single NVIDIA GeForce RTX 4090 GPU. The model is trained end to end with the ADAM optimizer and a batch size of 64.

The "*Drop Last*" issue is reported by several researchers (Qiu et al., 2024; 2025a; Li et al., 2025). That is, evaluating a model on the test set with drop-last set to True may cause additional errors related to the test batch size. In our experiments, to ensure a fair comparison, we set drop-last to False for all models to avoid this issue.

A.4 MLP Fusion Approach

For methods that do not support future covariates, we adapt them to incorporate future exogenous variables using an MLP fusion approach, as described in Algorithm 1. Note that, for fairness, all methods in our experiments use historical endogenous variables, historical exogenous variables, and future exogenous variables.

B Experimental Results

B.1 Visualization of Results

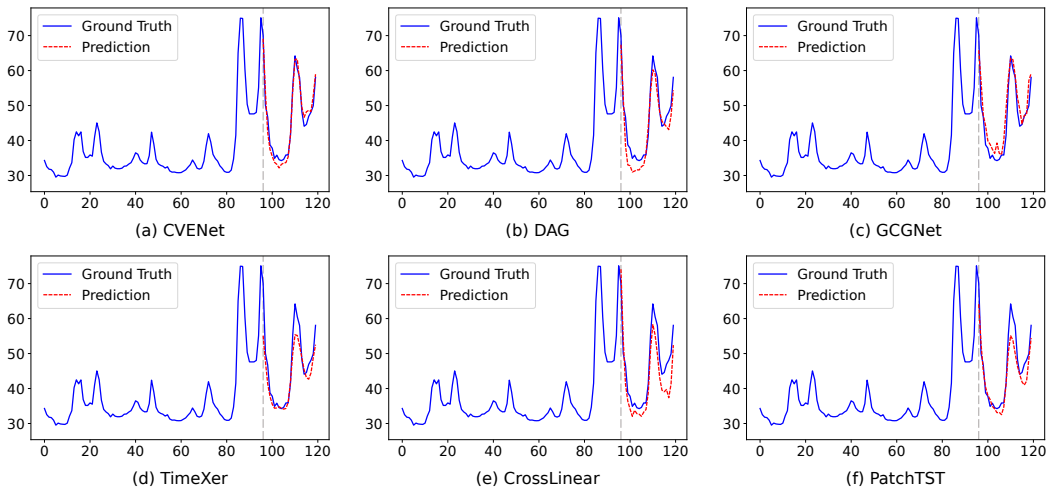


Figure 5: Visualization of prediction results on the NP dataset, where future exogenous variables are available.

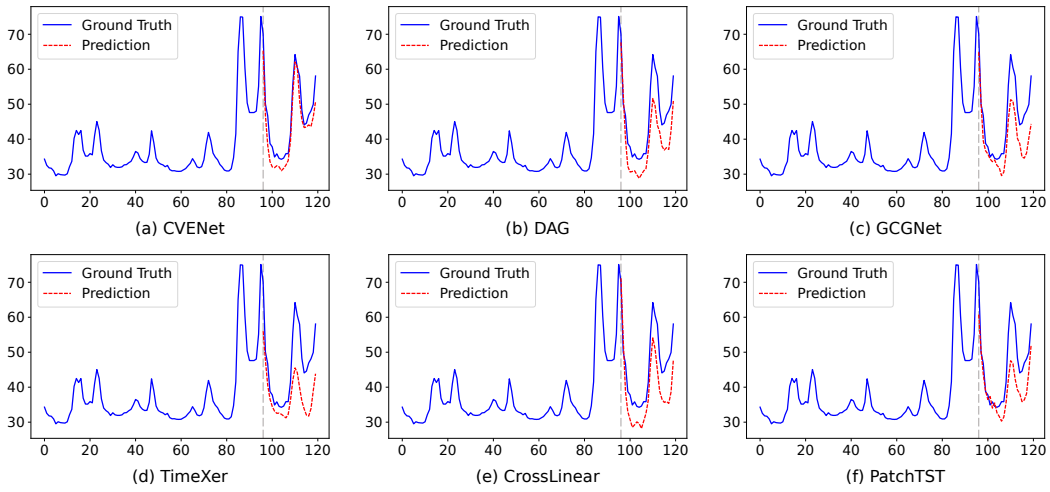


Figure 6: Visualization of prediction results on the NP dataset, where future exogenous variables are not available.

We visualize prediction cases on the NP dataset in Figure 5, where future exogenous variables are available, and in Figure 6, where they are not. Both figures show the same forecast window, so the two settings can be compared panel by panel. As shown in the two figures, CVENet follows the ground truth most closely in both settings, indicating that it makes fuller use of the exogenous

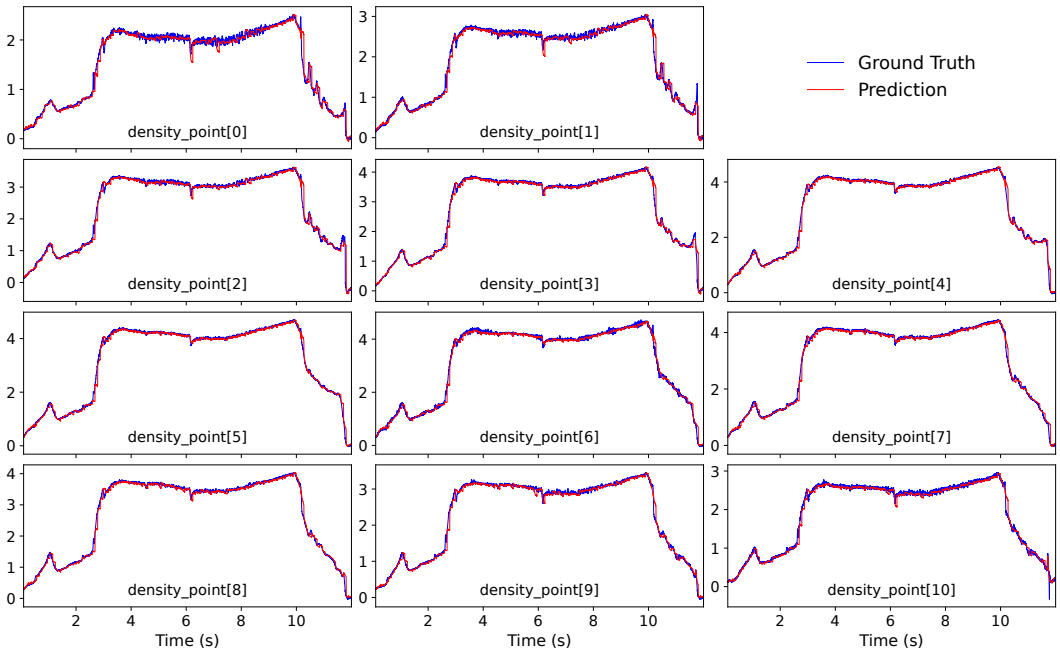


Figure 7: Prediction results throughout discharge #133361 of the EAST-POINT127 dataset.

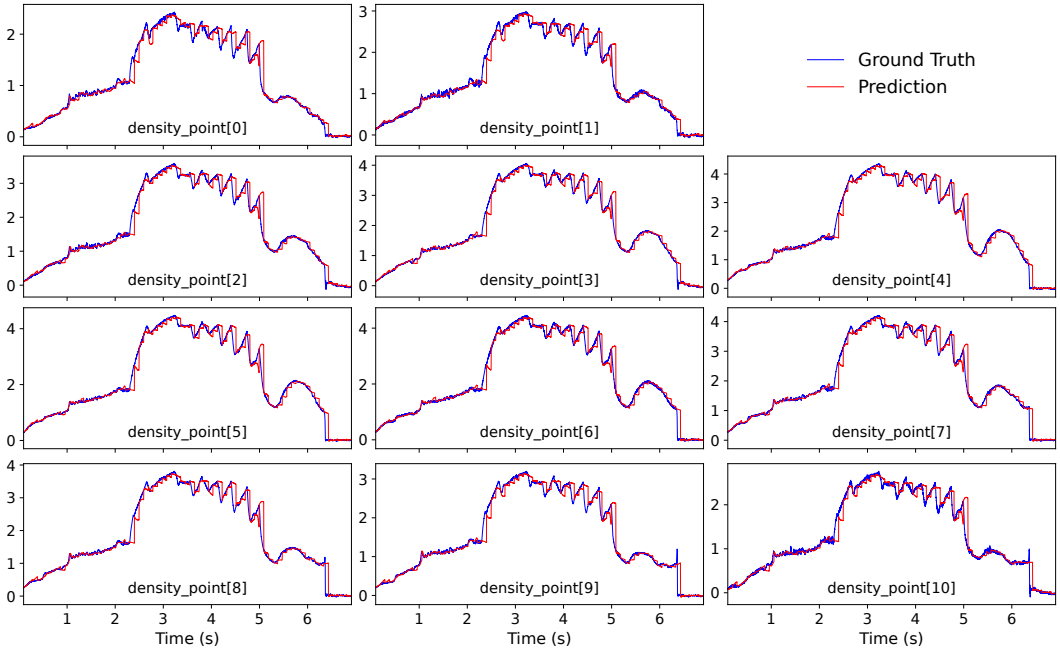


Figure 8: Prediction results throughout discharge #133089 of the EAST-POINT127 dataset.

variables than the compared methods. Withholding the future exogenous variables visibly degrades every method.

We further visualize the predictions of CVENet throughout an entire discharge of the EAST-POINT127 dataset in Figure 7, where the predictions are obtained by sliding the forecast window over the measured signals and concatenating the predictions (Ling et al., 2025). To analyze the failure cases, Figure 8 shows the test discharge on which CVENet has the highest error, about twice that of the discharge in Figure 7. Even on this discharge, the overall evolution of the density is

Table 7: Full results on 12 real-world datasets that satisfy the TSF-X task, where the inputs are (X^{endo}, X^{exo}, and Y^{exo}). **Red**: the best, Blue: the 2nd best. Avg means the average results from the two forecasting horizons.

Models		CVENet		DAG		GCGNet		TimeXer		TFT		TiDE		DUET		CrossLinear		Amplifier		TimeKAN		PatchTST	
Metrics		mse	mae	mse	mae	mse	mae	mse	mae	mse	mae	mse	mae	mse	mae	mse	mae	mse	mae	mse	mae	mse	mae
NP	24	**0.181**	**0.224**	0.202	0.237	0.208	0.237	0.236	0.266	0.219	0.249	0.284	0.301	0.246	0.287	0.210	0.266	0.252	0.303	0.273	0.310	0.249	0.294
	360	**0.503**	**0.436**	0.521	0.451	0.531	0.459	0.600	0.475	0.539	0.501	0.601	0.498	0.576	0.528	0.531	0.508	0.587	0.534	0.538	0.529	0.530	0.498
	Avg	**0.342**	**0.330**	0.362	0.344	0.370	0.348	0.418	0.371	0.379	0.375	0.443	0.400	0.411	0.408	0.371	0.387	0.420	0.418	0.405	0.419	0.390	0.396
PJM	24	**0.054**	**0.141**	0.057	0.143	0.060	0.150	0.075	0.166	0.095	0.195	0.106	0.214	0.072	0.166	0.088	0.191	0.096	0.208	0.115	0.244	0.116	0.239
	360	**0.107**	**0.210**	0.130	0.218	0.129	0.223	0.140	0.231	0.133	0.219	0.177	0.279	0.131	0.228	0.135	0.254	0.177	0.285	0.162	0.281	0.149	0.279
	Avg	**0.081**	**0.176**	0.093	0.180	0.095	0.187	0.108	0.198	0.114	0.207	0.142	0.246	0.102	0.197	0.112	0.223	0.137	0.246	0.139	0.262	0.133	0.259
BE	24	0.352	**0.221**	0.361	0.229	**0.350**	0.248	0.392	0.253	0.426	0.272	0.426	0.285	0.432	0.272	0.391	0.259	0.471	0.339	0.451	0.319	0.452	0.326
	360	0.524	0.358	0.485	0.330	0.511	0.340	0.512	0.327	**0.482**	**0.310**	0.571	0.364	0.597	0.436	0.568	0.416	0.646	0.487	0.645	0.495	0.702	0.538
	Avg	0.438	0.289	**0.423**	**0.279**	0.431	0.294	0.452	0.290	0.454	0.291	0.498	0.325	0.515	0.354	0.479	0.337	0.559	0.413	0.548	0.407	0.577	0.432
FR	24	0.348	**0.171**	0.355	0.171	**0.347**	0.188	0.366	0.208	0.543	0.253	0.418	0.255	0.384	0.251	0.390	0.226	0.459	0.348	0.454	0.296	0.518	0.368
	360	0.472	0.264	0.473	0.268	0.482	0.279	0.489	0.273	**0.465**	**0.261**	0.551	0.308	0.607	0.403	0.575	0.370	0.648	0.468	0.641	0.452	0.658	0.452
	Avg	**0.410**	**0.218**	0.414	0.219	0.415	0.234	0.427	0.241	0.504	0.257	0.484	0.281	0.496	0.327	0.483	0.298	0.554	0.408	0.547	0.374	0.588	0.410
DE	24	**0.271**	**0.310**	0.277	0.322	0.280	0.331	0.339	0.362	0.380	0.383	0.367	0.383	0.376	0.378	0.387	0.396	0.394	0.407	0.399	0.412	0.413	0.412
	360	0.473	0.421	**0.462**	**0.418**	0.523	0.447	0.610	0.474	0.599	0.509	0.630	0.511	0.589	0.482	0.583	0.507	0.551	0.474	0.547	0.479	0.589	0.498
	Avg	0.372	**0.365**	**0.370**	0.370	0.401	0.389	0.475	0.418	0.489	0.446	0.499	0.447	0.482	0.430	0.485	0.452	0.473	0.441	0.473	0.445	0.501	0.455
Energy	24	**0.070**	**0.202**	0.079	0.215	0.081	0.221	0.122	0.273	0.093	0.235	0.103	0.248	0.117	0.283	0.241	0.418	0.138	0.306	0.135	0.298	0.239	0.390
	360	0.183	0.334	0.169	**0.320**	0.182	0.334	0.204	0.357	**0.167**	0.331	0.202	0.355	0.288	0.452	0.237	0.385	0.328	0.472	0.302	0.464	0.214	0.363
	Avg	0.126	0.268	**0.124**	**0.267**	0.131	0.277	0.163	0.315	0.130	0.283	0.153	0.302	0.203	0.367	0.239	0.402	0.233	0.389	0.218	0.381	0.226	0.377
Sdwpfm1	24	**0.307**	**0.371**	0.351	0.400	0.376	0.415	0.558	0.533	0.366	0.421	0.474	0.488	0.551	0.564	0.355	0.473	0.364	0.445	0.418	0.503	0.374	0.454
	360	0.473	0.512	0.495	0.522	**0.472**	**0.499**	0.845	0.684	0.597	0.528	0.492	0.526	0.646	0.577	0.497	0.532	0.510	0.535	0.476	0.566	0.497	0.550
	Avg	**0.390**	**0.441**	0.423	0.461	0.424	0.457	0.701	0.609	0.482	0.474	0.483	0.507	0.599	0.570	0.426	0.502	0.437	0.490	0.447	0.534	0.435	0.502
Sdwpfm2	24	**0.340**	**0.391**	0.372	0.414	0.421	0.441	0.627	0.570	0.411	0.458	0.461	0.492	0.445	0.452	0.477	0.536	0.394	0.462	0.474	0.538	0.418	0.492
	360	0.521	0.542	0.583	0.556	0.529	0.531	0.978	0.736	0.541	**0.519**	**0.511**	0.540	0.584	0.528	0.589	0.611	0.587	0.563	0.520	0.589	0.602	0.603
	Avg	**0.430**	**0.466**	0.477	0.485	0.475	0.486	0.803	0.653	0.476	0.488	0.486	0.516	0.514	0.490	0.533	0.573	0.491	0.512	0.497	0.564	0.510	0.547
Sdwpfh1	24	**0.364**	**0.420**	0.408	0.438	0.435	0.486	0.651	0.587	0.401	0.460	0.434	0.489	0.527	0.513	0.548	0.585	0.576	0.627	0.511	0.582	0.422	0.505
	360	0.508	0.553	0.489	0.534	**0.465**	**0.514**	0.841	0.700	0.557	0.523	0.472	0.527	0.551	0.519	0.566	0.601	0.497	0.569	0.643	0.694	0.513	0.548
	Avg	**0.436**	**0.486**	0.448	0.486	0.450	0.500	0.746	0.643	0.479	0.491	0.453	0.508	0.539	0.516	0.557	0.593	0.537	0.598	0.577	0.638	0.468	0.527
Sdwpfh2	24	**0.407**	**0.452**	0.438	0.465	0.473	0.506	0.820	0.677	0.474	0.493	0.579	0.553	0.629	0.563	0.468	0.540	0.473	0.533	0.580	0.614	0.457	0.527
	360	0.598	0.589	0.608	0.595	**0.566**	0.565	0.962	0.761	0.657	**0.549**	0.619	0.614	0.665	0.569	0.608	0.609	0.569	0.628	0.713	0.729	0.641	0.632
	Avg	**0.503**	**0.520**	0.523	0.530	0.520	0.536	0.891	0.719	0.566	0.521	0.599	0.583	0.647	0.566	0.538	0.574	0.521	0.581	0.647	0.672	0.549	0.580
Colbun	10	**0.051**	**0.077**	0.061	0.094	0.065	0.108	0.113	0.172	0.092	0.135	0.089	0.131	0.089	0.134	0.071	0.102	0.071	0.121	0.061	0.101	0.062	0.122
	30	**0.127**	**0.179**	0.135	0.215	0.149	0.243	0.176	0.299	0.383	0.460	0.240	0.322	0.307	0.397	0.182	0.288	0.275	0.370	0.195	0.249	0.417	0.496
	Avg	**0.089**	**0.128**	0.098	0.154	0.107	0.175	0.145	0.235	0.238	0.297	0.164	0.227	0.198	0.266	0.126	0.195	0.173	0.246	0.128	0.175	0.239	0.309
Rapel	10	**0.147**	**0.200**	0.151	0.203	0.211	0.230	0.301	0.308	0.201	0.253	0.228	0.271	0.174	0.219	0.163	0.209	0.181	0.227	0.174	0.231	0.163	0.218
	30	**0.278**	**0.344**	0.309	0.408	0.401	0.384	0.387	0.416	0.409	0.414	0.411	0.432	0.365	0.432	0.340	0.417	0.333	0.416	0.325	0.390	0.374	0.445
	Avg	**0.212**	**0.272**	0.230	0.305	0.306	0.307	0.344	0.362	0.305	0.333	0.320	0.351	0.269	0.326	0.252	0.313	0.257	0.321	0.249	0.311	0.269	0.332
1st Count		**23**	**26**	4	4	5	2	0	0	3	4	1	0	0	0	0	0	0	0	0	0	0	0

followed, and the errors are concentrated at a few abrupt drops. Between 3.4 and 4.9 s, the density drops sharply several times while the actuator signals remain unchanged, so that neither the history nor the exogenous variables anticipate these drops, and the prediction catches up with each of them only in the next forecast window. The largest error occurs at about 5.0 s, when the density falls by about half within 0.1 s after one ECRH channel is switched off; although the switch-off is contained in the exogenous variables, the prediction remains at the previous level. Such responses depend on wall conditions and transport that are not among the measured variables (Wan et al., 2026), which limits what can be inferred from the actuator signals alone.

B.2 Full Results

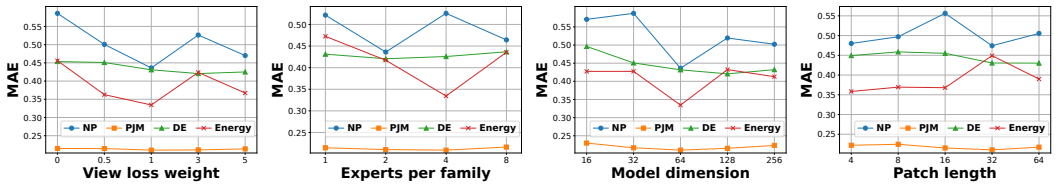
Full results on the 12 real-world datasets that satisfy the TSF-X task, where the inputs are X^{endo}, X^{exo}, and Y^{exo}, are reported in Table 7. Full results under the setting where only historical exogenous variables are used, where the inputs are X^{endo} and X^{exo}, are provided in Table 8.

B.3 Parameter Sensitivity

We also study the parameter sensitivity of CVENet, where all datasets use a lookback length of 720 and a forecasting horizon of 360 steps. We make the following observations: 1) Figure 9a illustrates the effect of the view loss weight γ. Switching the view loss off gives the largest error on every dataset, and the degradation is most pronounced on NP and Energy, highlighting the benefit of supervising each view as a predictor in its own right rather than only through the reconciled

Table 8: Full results under the setting where only historical exogenous variables are used, without future exogenous variables. The inputs are (X^{endo} and X^{exo}). **Red**: the best, Blue: the 2nd best. Avg means the average results from the two forecasting horizons.

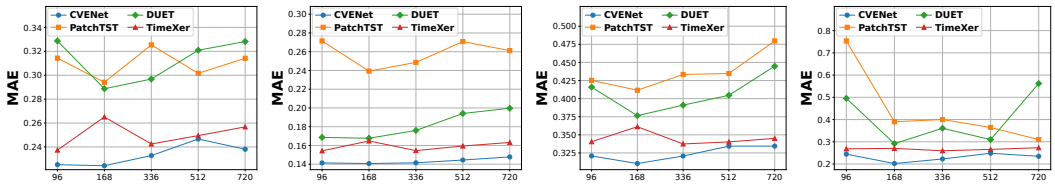
Models		CVENet		DAG		GCGNet		TimeXer		TFT		TiDE		DUET		CrossLinear		Amplifier		TimeKAN		PatchTST	
Metrics		mse	mae	mse	mae	mse	mae	mse	mae	mse	mae	mse	mae	mse	mae	mse	mae	mse	mae	mse	mae	mse	mae
NP	24	**0.219**	**0.255**	0.237	0.259	0.240	0.260	0.270	0.292	0.330	0.321	0.297	0.309	0.262	0.271	0.246	0.282	0.302	0.321	0.336	0.344	0.283	0.296
	360	0.645	0.512	**0.601**	0.500	0.620	0.502	0.609	**0.474**	0.964	0.655	0.638	0.523	0.626	0.494	0.655	0.505	0.738	0.574	0.632	0.527	0.630	0.505
	Avg	0.432	0.383	**0.419**	**0.380**	0.430	0.381	0.440	0.383	0.647	0.488	0.467	0.416	0.444	0.383	0.451	0.394	0.520	0.448	0.484	0.435	0.457	0.401
PJM	24	**0.072**	**0.160**	0.072	0.162	0.072	0.164	0.096	0.191	0.129	0.231	0.105	0.213	0.082	0.178	0.096	0.198	0.093	0.195	0.130	0.236	0.106	0.212
	360	**0.170**	**0.248**	0.180	0.259	0.191	0.269	0.186	0.267	0.270	0.310	0.210	0.292	0.198	0.273	0.197	0.283	0.210	0.294	0.219	0.304	0.189	0.273
	Avg	**0.121**	**0.204**	0.126	0.210	0.131	0.216	0.141	0.229	0.200	0.270	0.158	0.253	0.140	0.226	0.147	0.241	0.152	0.245	0.175	0.270	0.148	0.243
BE	24	0.369	**0.237**	0.376	0.244	**0.365**	0.259	0.392	0.253	0.400	0.267	0.492	0.322	0.376	0.253	0.389	0.253	0.417	0.283	0.424	0.280	0.405	0.264
	360	0.555	0.377	**0.547**	0.350	0.553	0.354	0.563	0.349	0.727	0.442	0.602	0.375	0.571	0.357	0.564	**0.346**	0.588	0.367	0.552	0.350	0.564	0.354
	Avg	0.462	0.307	0.462	**0.297**	**0.459**	0.306	0.477	0.301	0.563	0.354	0.547	0.348	0.473	0.305	0.477	0.300	0.502	0.325	0.488	0.315	0.485	0.309
FR	24	0.354	**0.189**	**0.345**	0.190	0.352	0.199	0.366	0.195	0.445	0.231	0.415	0.254	0.359	0.208	0.397	0.204	0.429	0.250	0.448	0.250	0.397	0.223
	360	0.542	0.327	0.526	0.307	**0.519**	**0.293**	0.542	0.300	0.624	0.345	0.574	0.325	0.577	0.317	0.555	0.310	0.559	0.322	0.534	0.303	0.542	0.306
	Avg	0.448	0.258	**0.435**	0.249	0.435	**0.246**	0.454	0.247	0.535	0.288	0.494	0.290	0.468	0.262	0.476	0.257	0.494	0.286	0.491	0.276	0.470	0.264
DE	24	**0.412**	**0.395**	0.422	0.411	0.413	0.402	0.501	0.445	0.576	0.452	0.465	0.433	0.424	0.403	0.438	0.423	0.493	0.455	0.504	0.462	0.503	0.450
	360	**0.772**	**0.561**	0.784	0.563	0.816	0.600	0.818	0.568	0.792	0.578	0.823	0.604	0.895	0.623	0.832	0.592	0.930	0.642	0.834	0.590	0.889	0.604
	Avg	**0.592**	**0.478**	0.603	0.487	0.614	0.501	0.659	0.507	0.684	0.515	0.644	0.519	0.660	0.513	0.635	0.508	0.712	0.548	0.669	0.526	0.696	0.527
Energy	24	0.111	0.255	0.112	0.257	0.112	0.259	0.138	0.293	0.352	0.463	0.117	0.265	0.108	0.254	**0.102**	**0.246**	0.108	0.254	0.109	0.254	0.108	0.254
	360	0.196	0.349	0.189	0.342	0.195	0.351	0.206	0.360	0.399	0.497	0.207	0.358	0.211	0.362	0.299	0.425	0.251	0.400	**0.186**	**0.338**	0.298	0.428
	Avg	0.154	0.302	0.150	0.300	0.154	0.305	0.172	0.326	0.376	0.480	0.162	0.311	0.160	0.308	0.201	0.336	0.180	0.327	**0.147**	**0.296**	0.203	0.341
Sdwpfm1	24	**0.508**	**0.473**	0.532	0.512	0.537	0.509	0.558	0.533	0.596	0.531	0.561	0.540	0.550	0.495	0.549	0.553	0.579	0.557	0.598	0.579	0.557	0.557
	360	**0.844**	0.684	0.845	0.695	0.885	0.682	0.845	0.684	1.372	0.925	0.864	0.725	0.898	**0.672**	1.070	0.834	1.106	0.814	0.853	0.764	0.910	0.745
	Avg	**0.676**	**0.578**	0.689	0.604	0.711	0.595	0.701	0.609	0.984	0.728	0.713	0.633	0.724	0.583	0.809	0.694	0.843	0.685	0.725	0.672	0.733	0.651
Sdwpfm2	24	**0.583**	**0.518**	0.609	0.548	0.630	0.547	0.627	0.570	0.749	0.657	0.636	0.582	0.612	0.537	0.635	0.586	0.669	0.600	0.684	0.621	0.648	0.610
	360	0.963	**0.723**	**0.962**	0.766	0.983	0.739	0.978	0.736	1.340	0.877	0.995	0.783	1.028	0.745	0.965	0.788	1.215	0.863	0.987	0.780	1.020	0.819
	Avg	**0.773**	**0.620**	0.786	0.657	0.807	0.643	0.803	0.653	1.044	0.767	0.816	0.682	0.820	0.641	0.800	0.687	0.942	0.732	0.836	0.701	0.834	0.714
Sdwpfh1	24	**0.636**	**0.572**	0.647	0.592	0.651	0.578	0.651	0.587	0.754	0.670	0.719	0.641	0.673	0.575	0.687	0.652	1.111	0.816	0.755	0.679	0.669	0.629
	360	0.827	0.686	**0.820**	0.695	0.830	**0.677**	0.841	0.700	0.831	0.727	0.897	0.757	0.886	0.712	0.849	0.752	1.089	0.823	0.853	0.762	0.939	0.804
	Avg	**0.731**	0.629	0.733	0.643	0.741	**0.628**	0.746	0.643	0.793	0.698	0.808	0.699	0.779	0.644	0.768	0.702	1.100	0.820	0.804	0.720	0.804	0.717
Sdwpfh2	24	**0.719**	**0.622**	0.739	0.635	0.798	0.645	0.820	0.677	0.811	0.714	0.783	0.674	0.975	0.694	0.810	0.718	0.834	0.691	0.892	0.740	0.769	0.680
	360	0.980	0.746	1.000	0.755	0.974	0.757	**0.962**	0.761	1.042	0.779	1.056	0.829	1.039	**0.735**	1.102	0.830	1.126	0.845	0.990	0.825	1.280	0.925
	Avg	**0.849**	**0.684**	0.870	0.695	0.886	0.701	0.891	0.719	0.926	0.746	0.919	0.751	1.007	0.715	0.956	0.774	0.980	0.768	0.941	0.782	1.025	0.802
Colbun	10	0.071	**0.094**	0.072	0.103	0.070	0.094	0.084	0.146	0.449	0.279	0.095	0.140	0.073	0.106	**0.068**	0.101	0.077	0.120	0.091	0.129	0.083	0.125
	30	0.173	0.215	**0.161**	**0.208**	0.168	0.236	0.181	0.293	0.663	0.494	0.282	0.341	0.221	0.290	0.190	0.289	0.227	0.301	0.311	0.377	0.217	0.316
	Avg	0.122	**0.155**	**0.117**	0.155	0.119	0.165	0.132	0.219	0.556	0.386	0.188	0.240	0.147	0.198	0.129	0.195	0.152	0.210	0.201	0.253	0.150	0.221
Rapel	10	**0.195**	**0.214**	0.195	0.215	0.217	0.221	0.301	0.308	0.238	0.274	0.238	0.270	0.221	0.225	0.201	0.224	0.240	0.259	0.226	0.238	0.215	0.232
	30	0.281	**0.339**	0.327	0.347	0.313	0.346	0.387	0.416	0.377	0.394	0.409	0.444	0.387	0.386	**0.278**	0.374	0.433	0.436	0.458	0.437	0.436	0.436
	Avg	**0.238**	**0.277**	0.261	0.281	0.265	0.284	0.344	0.362	0.308	0.334	0.323	0.357	0.304	0.306	0.240	0.299	0.337	0.348	0.342	0.337	0.325	0.334
1st Count		**18**	**22**	9	3	3	4	1	1	0	0	0	0	0	2	3	2	0	0	2	2	0	0



(a) View loss weight (b) Experts per family (c) Model dimension (d) Patch length

Figure 9: Parameter sensitivity studies of main hyper-parameters in CVENet.

forecast. Once the loss is active, a weight of 1 yields the best results in most cases, indicating that moderate weighting is beneficial. 2) Figure 9b illustrates the effect of the number of experts per family $E/2$. Assigning a single expert to each family is never optimal on any dataset, indicating that specialization within a family contributes beyond the choice between the two functional forms themselves. Two to four experts per family yield the best results, whereas a larger pool brings no further gain and only increases the parameter count. 3) Figure 9c shows that the model dimension d has a certain impact on performance, but CVENet maintains good stability across different configurations. Larger dimensions do not always improve forecasting accuracy, since excessive capacity introduces redundancy and makes optimization more difficult, while dimensions in the range of 64 to 128 generally achieve a good balance between prediction accuracy and model complexity. 4) Figure 9d explores the effect of patch length on model performance. Due to differences in temporal



(a) NP (b) PJM (c) DE (d) Energy

Figure 10: Forecasting performance (MAE) with varying look-back windows on NP, PJM, DE and Energy. The look-back windows are 96, 168, 336, 512 and 720, and the forecasting horizon is 24.

dependencies across datasets, the optimal patch length varies; however, a patch length of 32 tends to yield favorable results in most cases. It is worth noting that very small patch lengths increase the number of tokens and therefore the computational cost, while excessively large ones may weaken the ability of the model to capture local dependencies.

B.4 Increasing Look-back Windows

The size of the look-back window determines how much history a forecaster receives. A longer window supplies more information, but it also carries more noise, and the effect is amplified in forecasting with exogenous variables, where enlarging the window lengthens every covariate series as well. We therefore configure look-back windows of 96, 168, 336, 512 and 720 on NP, PJM, DE and Energy, with the forecasting horizon fixed at 24, and compare CVENet against three representative baselines. As shown in Figure 10, CVENet attains the lowest error at every look-back window on all four datasets. The two baselines that do not natively consume future exogenous variables fluctuate the most as the window grows, whereas the error of CVENet stays within a narrow band on every dataset. On PJM, DE and Energy the margin is wide enough that the worst result of CVENet over the whole range is still better than the best result any baseline attains at any look-back window, demonstrating the robustness of CVENet across different amounts of historical context.

B.5 Limitations and Future Works

Despite the consistent improvements reported above, CVENet still has several limitations. First, the two complementary views are carried by two stacks of encoder layers that share no parameters, so the parameters and the computation grow with the number of views, making a richer set of statistical views more costly to explore. Sharing parameters across views, or deciding per series which views are worth running at all, is left to future work. Second, the two settings considered in this work are the two extremes of covariate availability: either every exogenous variable is known over the forecast horizon, or none of them is. Deployments often fall in between, with some covariates known in advance and others not, and extending the model to arbitrary subsets of known covariates is an important direction for future work.