MIXROUTE: MIXED PARAMETERIZATION WITH ADAPTIVE INFORMATION ROUTING FOR TIME SERIES FORECASTING WITH EXOGENOUS VARIABLES

Anonymous authors
Paper under double-blind review

ABSTRACT

Time series forecasting with exogenous variables (TSF-X) predicts an endogenous target series from its own history and from exogenous variables whose future values may be available in advance. The task therefore couples temporal dependence within the endogenous series with cross-variable dependence on the exogenous variables. Decoupled methods model these dependencies in separate modules and rely on their combination to capture correlations that span time and variables jointly. Joint methods instead model both dependencies in one token space, but parameter sharing and information routing remain important design considerations. Homogeneous parameterization applies the same transformations to endogenous patches and exogenous-variable tokens despite their different semantic roles. A single attention distribution per head couples temporal modeling with exogenous information acquisition, without explicit specialization for either role. To address these issues, we propose MixRoute, a unified Transformer framework that makes representation and information flow type-aware for endogenous patches, exogenous variables, and learnable global tokens. Under Mixed Parameterization, endogenous patches share one set of parameters, whereas each exogenous and global token has a private set. Adaptive Information Routing (AIR) uses separate queries for an intrinsic stream over the endogenous history and a cross-type stream over the exogenous and global tokens. It then combines the stream outputs with a gate computed from both streams. On 12 real-world datasets in the with-future setting, MixRoute reduces MSE by 8.0% on average relative to the strongest baseline evaluated. When future exogenous values are unavailable, MixRoute needs no architectural change and still ranks first overall.

1 INTRODUCTION

Time series forecasting is essential in many real-world applications, including energy management (Zhou et al., 2025b; Niu et al., 2025), traffic planning (Shao et al., 2026; Zhang et al., 2026a), financial analysis (Bao et al., 2025), healthcare monitoring (Yu et al., 2026) and industrial systems (Ren et al., 2023). Many univariate and multivariate forecasting methods rely solely on historical observations of target variables (Nie et al., 2023; Liu et al., 2024; Zhao et al., 2024a;b). In practice, exogenous variables (i.e., covariates) can provide additional information about the target. In energy demand forecasting, for instance, temperature and other weather conditions strongly shape consumption patterns (Qin et al., 2023; Dong et al., 2025). Forecasts of these weather conditions can be available before the target prediction is made. In retail sales prediction, holiday schedules and promotional calendars are exogenous variables that shape demand fluctuations (Chen et al., 2025). In this work, we focus on time series forecasting with exogenous variables (TSF-X). This task uses historical exogenous values to help predict endogenous target variables and incorporates future exogenous values when available. A TSF-X model must capture two kinds of dependency at once. Temporal correlations characterize how the endogenous series evolves from its own past. Cross-variable correlations characterize its dependence on exogenous variables. How these two are organized within a single architecture is the central design question in TSF-X.

Recent TSF-X methods fall into two broad families, distinguished by how they answer this question. Figure 1 contrasts the two families with MixRoute, and Appendix B reviews prior work in detail.

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.

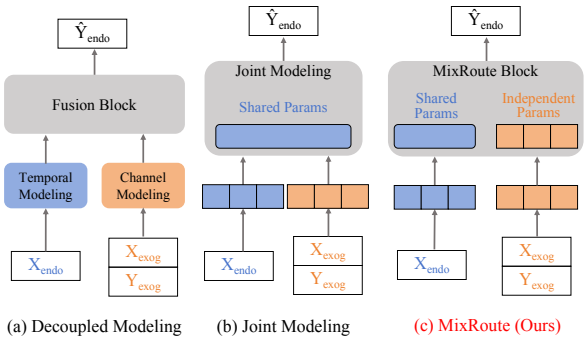


Figure 1: Categorization of TSF-X methods. (a) **Decoupled**: separate modules model temporal and cross-variable (channel) correlations. (b) **Joint modeling**: one block models both with shared parameters. (c) **MixRoute**: joint modeling with Mixed Parameterization; endogenous patches share parameters, while each exogenous variable has a private set.

Decoupled Modeling approaches model temporal and cross-variable correlations in separate modules. Most arrange the two in sequence. Some capture temporal correlations first, then fuse cross-variable information via cross-attention (Wang et al., 2024; Tayal et al., 2024). Others start from cross-variable mixing and model temporal correlations afterwards (Zhou et al., 2025a; Lim et al., 2021). DAG (Qiu et al., 2026) instead runs the two in parallel. These approaches therefore rely on module composition to capture dependencies spanning time and variables (Li et al., 2026). Sequential arrangements may also introduce interference between stages. Temporal-first encoding may produce representations suboptimal for subsequent cross-variable interaction, whereas cross-variable-first mixing may distort fine-grained endogenous temporal patterns before they are captured.

Joint Modeling approaches such as GCGNet (Li et al., 2026) model temporal and cross-variable correlations together in a unified token space. When endogenous history is represented as patches and each exogenous variable as one token, homogeneous parameterization shares transformations across tokens with different semantic roles. Endogenous patches are consecutive segments of the same variable and have a natural temporal ordering. Exogenous representations, however, are heterogeneous. They have no inherent ordering among them, and each represents a distinct variable with its own distributional characteristics (Liu et al., 2024; Zhang et al., 2026b). Wind speed, temperature and surface pressure at a wind-power site differ in scale, dynamics, and their relationships with the target. Sharing parameters across these two types assumes that one transformation can extract both the endogenous temporal patterns and the exogenous driving signals.

Beyond shared parameterization, joint modeling leaves open how an endogenous token should divide attention between its own history and the exogenous variables. In attention-based joint models, each endogenous token uses a single query per head to attend to both token types. The attention scores are normalized together, so the same scores determine both how attention is distributed within each type and how much total attention each type receives. Temporal modeling and exogenous information acquisition have different structural requirements. The former is order-sensitive, whereas the latter operates over unordered variables. Although different heads may learn to specialize in either task, the architecture does not explicitly assign these roles. We refer to this lack of explicit specialization as entangled attention.

The two limitations concern complementary aspects of joint modeling. Homogeneous parameterization concerns how heterogeneous tokens are represented, whereas entangled attention concerns how information flows among them. With type-aware parameterization alone, each endogenous query still uses one attention distribution for both temporal modeling and exogenous information acquisition. Conversely, separating the attention streams does not determine which tokens share parameters. Type-aware parameterization has been studied in recommender systems, including HiFormer (Gui et al., 2023) and OneTrans (Zhang et al., 2026b). In particular, OneTrans shares attention projections and FFNs across sequential tokens while assigning token-specific parameters to non-sequential tokens. MixRoute applies the same parameter-sharing principle to endogenous patches and exogenous/global tokens. Under OneTrans's causal token ordering, sequential tokens cannot attend to non-sequential tokens. In contrast, AIR allows each endogenous patch to attend to exogenous and global tokens through a separate cross-type stream.

Based on the above observations, we propose **MixRoute** (**Mix**ed parameterization with adaptive information **Rout**ing), a unified Transformer framework that addresses both limitations through two mechanisms. MixRoute processes three semantic token types: endogenous patches, exogenous variables, and learnable global tokens that serve as cross-type information hubs. Each block combines

Mixed Attention, Mixed FFN and AIR. Mixed Attention and Mixed FFN share parameters across endogenous patches and use private parameters for each exogenous and global token. The Adaptive Information Routing (AIR) mechanism separates attention for endogenous tokens into two streams. An intrinsic stream captures temporal dependence within the series, and a cross-type stream acquires exogenous information. A learned, input-dependent gate combines the stream outputs. The two streams use separate endogenous queries while reusing the keys and values constructed under Mixed Parameterization. Heterogeneous tokens are therefore modeled end to end, with no separate fusion stage after the encoder. Our contributions are summarized as follows:

- We develop MixRoute, a unified Transformer framework that makes representation and information flow type-aware for TSF-X. Mixed Parameterization shares attention projections and FFN parameters across endogenous patches while assigning private parameters to each exogenous and global token.
- We introduce Adaptive Information Routing, which separates within-group information selection from stream combination. The intrinsic and cross-type streams use separate endogenous queries and separately normalized attention, with an input-dependent gate combining their outputs. The pooled stream outputs are regularized toward orthogonality.
- In the with-future setting, extensive experiments on 12 real-world TSF-X datasets show that MixRoute reduces MSE by 8.0% on average relative to the strongest baseline, DAG, and by 11.9% relative to GCGNet. When future exogenous values are unavailable, MixRoute needs no architectural change and still ranks first overall.

2 METHODOLOGY

2.1 PROBLEM FORMULATION

Given the historical endogenous series $\mathbf{X}_{\text{endo}} \in \mathbb{R}^{N\times T}$, the historical exogenous series $\mathbf{X}_{\text{exog}} \in \mathbb{R}^{D\times T}$, and the future exogenous series $\mathbf{Y}_{\text{exog}} \in \mathbb{R}^{D\times F}$, the goal of TSF-X is to predict the future endogenous values:

$$\hat{\mathbf{Y}}_{\text{endo}} = \mathcal{F}_\theta(\mathbf{X}_{\text{endo}}, \mathbf{X}_{\text{exog}}, \mathbf{Y}_{\text{exog}}) \tag{1}$$

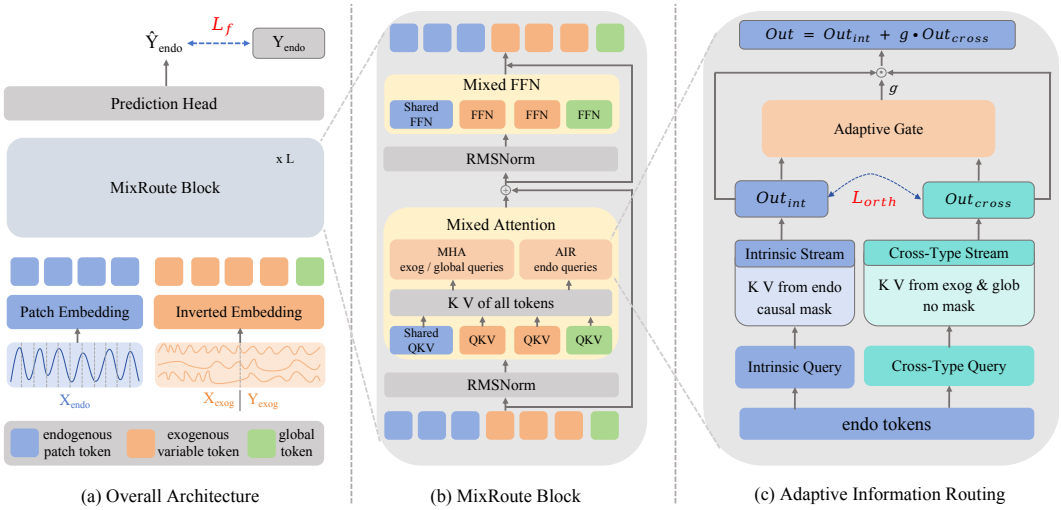
where $\hat{\mathbf{Y}}_{\text{endo}} \in \mathbb{R}^{N\times F}$ and $\mathcal{F}_\theta$ denotes the forecasting model parameterized by θ. Here N is the number of target variables, D the number of exogenous variables, T the number of past time steps, and F the forecasting horizon. We focus on the single-target case ($N = 1$), as in our experiments. Two settings arise in practice. In the with-future setting, $\mathbf{Y}_{\text{exog}}$ is available and Equation (1) applies as written. In the without-future setting, the forecast must be produced from $(\mathbf{X}_{\text{endo}}, \mathbf{X}_{\text{exog}})$ alone.

2.2 STRUCTURE OVERVIEW

The overall architecture of MixRoute is illustrated in Figure 2. MixRoute consists of three stages: (1) heterogeneous token embedding (Section 2.3), (2) a stack of L MixRoute Blocks (Sections 2.4 and 2.5), and (3) a prediction head (Section 2.6). The embedding stage maps the endogenous history to patch tokens and each exogenous variable to one token, then adds a set of learnable global tokens. The three types are concatenated into a single sequence. **Mixed Parameterization** enforces one rule from the embedding through every block: endogenous tokens share one set of parameters, whereas each exogenous and global token has a private set. **Adaptive Information Routing** operates within the attention sublayer of each block and applies only to endogenous queries. Exogenous information reaches the forecast along two paths. The direct path is AIR's cross-type stream. The indirect path runs through the global tokens. They accumulate information from both token types in every block, and the prediction head maps them to the forecast.

2.3 HETEROGENEOUS TOKEN EMBEDDING

This stage produces three token types: endo-tokens from the endogenous history, exog-tokens from the exogenous variables, and learnable global tokens. We apply instance normalization to the inputs and de-normalize the forecast at the output. For the endogenous series we use patch-wise tokenization: $\mathbf{X}_{\text{endo}}$ is split into $N_e = \lfloor T/P \rfloor$ non-overlapping patches of length P, and each patch is



(a) Overall Architecture (b) MixRoute Block (c) Adaptive Information Routing

Figure 2: Overall architecture of MixRoute. (a) Heterogeneous token embedding and the overall pipeline. (b) The MixRoute Block: Mixed Attention and Mixed FFN use shared parameters for endogenous tokens and private parameters for each exogenous or global token. (c) Adaptive Information Routing: a learned gate $\mathbf{g}$ combines the outputs of the intrinsic and cross-type streams.

linearly projected to the model width d:

$$\mathbf{E}_{\text{endo}} = \text{PatchEmbed}(\mathbf{X}_{\text{endo}}) \in \mathbb{R}^{N_e \times d}. \tag{2}$$

Each exogenous variable is embedded as a single token by inverted embedding (Liu et al., 2024). The historical and future segments of each exogenous variable are concatenated along the temporal dimension, and every variable has a private projection matrix:

$$\begin{aligned} \mathbf{e}^{(i)}_{\text{exog}} &= \left[\mathbf{x}^{(i)}_{\text{exog}}, \mathbf{y}^{(i)}_{\text{exog}}\right] \mathbf{W}^{(i)}_{\text{exog}} \in \mathbb{R}^{d}, \\ \mathbf{E}_{\text{exog}} &= \left[\mathbf{e}^{(1)}_{\text{exog}}; \ldots; \mathbf{e}^{(D)}_{\text{exog}}\right] \in \mathbb{R}^{D \times d} \end{aligned} \tag{3}$$

with $\mathbf{W}^{(i)}_{\text{exog}} \in \mathbb{R}^{(T_x+F_x)\times d}$, where T_x and F_x count the retained historical and future exogenous observations. In the aligned setting, $T_x = T_e = PN_e \le T$ and $F_x = F$; both histories use the most recent T_e steps. No positional encoding is applied to exog-tokens, whose index runs over variables rather than time steps. Without future inputs, $F_x = 0$ and $\mathbf{e}^{(i)}_{\text{exog}} = \mathbf{x}^{(i)}_{\text{exog}}\mathbf{W}^{(i)}_{\text{exog}}$, with $\mathbf{W}^{(i)}_{\text{exog}} \in \mathbb{R}^{T_x \times d}$. Nothing else in the architecture changes.

The N_g global tokens $\mathbf{E}_{\text{glob}} \in \mathbb{R}^{N_g \times d}$ are randomly initialized learnable parameters. Unlike the endogenous patches, they receive no positional encoding and are not causally masked. In every block they attend to both endo- and exog-tokens. This makes each global token a compact summary of the whole sequence. The main branch of the prediction head later reads out this summary. To distinguish the three token types, we add a learnable type embedding $\mathbf{E}_{\text{type}}$ to the concatenated sequence:

$$\mathbf{Z}^{(0)} = [\mathbf{E}_{\text{endo}}; \mathbf{E}_{\text{exog}}; \mathbf{E}_{\text{glob}}] + \mathbf{E}_{\text{type}} \in \mathbb{R}^{M \times d} \tag{4}$$

where $M = N_e + D + N_g$ is the total number of tokens and $\mathbf{Z}^{(0)}$ denotes the input to the first MixRoute Block.

2.4 Mixed Parameterization

Each MixRoute Block processes the heterogeneous token sequence through Mixed Attention and Mixed FFN, both with pre-normalization and residual connections. We write $\mathbf{Z}^{(l)}$ for the output of block l. A subscript s marks quantities of the endo-tokens, which form an ordered sequence of patches. A subscript u marks those of the exog- and global tokens, which form an unordered set.

Mixed Attention. Let $\mathbf{H} = \text{RMSNorm}(\mathbf{Z}^{(l-1)})$ be split into the N_e endo-tokens $\mathbf{H}_s$ and the $D+N_g$ exog- and global tokens. Every token is projected into query, key and value vectors. Within attention,

Mixed Parameterization specifies which tokens share the projection weights. All endo-tokens share the query, key and value projection matrices, whereas each exogenous and global token has a private set of the same shapes. Each projection matrix has shape $d \times d$. Only the endo-token queries and keys receive RoPE (Su et al., 2024), so their attention scores depend on the relative offset between patches. The exog- and global tokens receive no positional encoding. The two groups are stacked into $\mathbf{Q}, \mathbf{K}, \mathbf{V} \in \mathbb{R}^{M \times d}$, and a single attention is computed jointly over all M tokens:

$$\text{MixedAttn}(\mathbf{H}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d_h}}\right)\mathbf{V}\mathbf{W}_O \tag{5}$$

with H heads of dimension $d_h = d/H$ and a shared output projection $\mathbf{W}_O$. For brevity, the attention equations omit the head index; attention is computed separately within each head, and the head outputs are concatenated. Equation (5) is the base form of the block: AIR (Section 2.5) replaces the output rows of the endo-token queries and leaves the exog- and global rows unchanged.

Mixed FFN. The feed-forward sublayer follows the same rule: all endo-tokens share one network, and each exogenous and global token has a private one. Each is a gated network

$$\text{FFN}(\mathbf{x}) = \big(\text{GELU}(\mathbf{x}\mathbf{W}^{\text{gate}}) \odot \mathbf{x}\mathbf{W}^{\text{up}}\big)\mathbf{W}^{\text{down}} \tag{6}$$

where $\mathbf{W}^{\text{gate}}, \mathbf{W}^{\text{up}} \in \mathbb{R}^{d \times d_{\text{ff}}}$, $\mathbf{W}^{\text{down}} \in \mathbb{R}^{d_{\text{ff}} \times d}$, and $\odot$ is element-wise multiplication. The per-token networks differ only in weights and are evaluated in parallel using batched matrix multiplications.

2.5 Adaptive Information Routing

Under Mixed Parameterization, each endo-token still uses one query and one jointly normalized attention distribution per head. **Adaptive Information Routing** uses separate endogenous queries for an intrinsic stream over endo-tokens and a cross-type stream over exog- and global tokens. The two attention distributions are normalized separately, rather than jointly over both token groups. An input-dependent gate scales the cross-type output, which is then added to the intrinsic output. Both streams reuse the keys and values of Mixed Attention.

Intrinsic Stream. The intrinsic stream applies a dedicated query projection $\mathbf{W}^Q_{\text{int}}$ with RoPE and attends only to the endo keys and values under a causal mask:

$$\mathbf{O}^{\text{int}} = \text{softmax}\left(\frac{\text{RoPE}(\mathbf{H}_s\mathbf{W}^Q_{\text{int}})\,\tilde{\mathbf{K}}_s^\top}{\sqrt{d_h}} + \mathbf{M}_{\text{causal}}\right)\mathbf{V}_s \tag{7}$$

where $\mathbf{W}^Q_{\text{int}} \in \mathbb{R}^{d \times d}$, $\tilde{\mathbf{K}}_s$ and $\mathbf{V}_s$ are the RoPE-rotated endo keys and the endo values of Mixed Attention, $\mathbf{M}_{\text{causal}}$ is the additive causal mask, and $\mathbf{O}^{\text{int}} \in \mathbb{R}^{N_e \times d}$ is concatenated over heads.

Cross-Type Stream. The cross-type stream uses a separate query projection $\mathbf{W}^Q_{\text{cross}}$ of the same shape without RoPE and attends to the exog- and global keys and values $\mathbf{K}_u, \mathbf{V}_u$ with no mask:

$$\mathbf{O}^{\text{cross}} = \text{softmax}\left(\frac{(\mathbf{H}_s\mathbf{W}^Q_{\text{cross}})\,\mathbf{K}_u^\top}{\sqrt{d_h}}\right)\mathbf{V}_u \tag{8}$$

The two streams therefore have distinct structural roles. RoPE and the causal mask endow the intrinsic stream with an order-aware temporal inductive bias, while the cross-type stream treats exogenous variables as an unordered set.

Gated Stream Fusion. An input-dependent gate scales the cross-type stream output for each head and sample. The gate fuses the two streams into the endogenous output:

$$\mathbf{g} = \text{sigmoid}\big(\text{LN}\big(\big[\mathbf{o}^{\text{int}}, \mathbf{o}^{\text{cross}}\big]\big)\,\mathbf{W}_g + \mathbf{b}_g\big), \quad \mathbf{O}_s = \mathbf{O}^{\text{int}} + \mathbf{g} \odot \mathbf{O}^{\text{cross}}, \tag{9}$$

where $\mathbf{o}^{\text{int}}, \mathbf{o}^{\text{cross}} \in \mathbb{R}^d$ are obtained by average pooling the corresponding stream outputs over the endo-token dimension and $\mathbf{g} \in \mathbb{R}^H$. The h-th entry of $\mathbf{g}$ scales the h-th head block of $\mathbf{O}^{\text{cross}}$. Here, LN denotes layer normalization and $\mathbf{b}_g \in \mathbb{R}^H$ is a learnable bias. Exog- and global tokens are updated by standard attention over all M tokens. All M rows then pass through the shared output projection $\mathbf{W}_O$, unchanged from Mixed Attention.

Table 1: Results on 12 real-world TSF-X datasets, where the inputs are ($\mathbf{X}_{\text{endo}}$, $\mathbf{X}_{\text{exog}}$ and $\mathbf{Y}_{\text{exog}}$). All values are averaged over the two forecasting horizons. Red: the best, Blue: the 2nd best. Full per-horizon results: Table 5 in Appendix C.3.

Models	**MixRoute**		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	**0.289**	**0.308**	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	**0.090**	0.184	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	**0.399**	**0.265**	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	**0.369**	**0.210**	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	**0.347**	**0.362**	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	**0.109**	**0.252**	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	**0.416**	**0.457**	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	**0.465**	0.489	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	**0.432**	**0.479**	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	**0.496**	**0.517**	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	**0.087**	**0.145**	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	**0.198**	**0.260**	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	**12**	**10**	0	2	0	1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

2.6 Prediction and Training Objective

Prediction Head. Global tokens serve a second function at the end of the encoder: they are the input to the prediction head. Let $\mathbf{Z}_{\text{glob}}^{(L)} \in \mathbb{R}^{N_g \times d}$ and $\mathbf{Z}_{\text{endo}}^{(L)} \in \mathbb{R}^{N_e \times d}$ be the final-layer global and endo-token states. The main branch flattens the global tokens and maps the resulting vector to the forecast with an MLP. Because global tokens attend to endo- and exog-tokens at every layer, this input already integrates both token types, but it compresses the whole sequence into N_g tokens. A second lightweight branch bypasses this bottleneck with a linear shortcut that maps the endo-tokens straight to the forecast:

$$\hat{\mathbf{Y}}_{\text{endo}} = \text{MLP}\Big(\text{flatten}\big(\mathbf{Z}_{\text{glob}}^{(L)}\big)\Big) + \gamma \left(\mathbf{W}_t \, \mathbf{Z}_{\text{endo}}^{(L)} \, \mathbf{W}_c\right)^{\top} \tag{10}$$

where $\mathbf{W}_t \in \mathbb{R}^{F \times N_e}$ maps the N_e patches to the F future steps, $\mathbf{W}_c \in \mathbb{R}^{d \times N}$ maps the features to the N target channels, and the learnable scalar γ scales the shortcut.

Training Objective. The primary objective $\mathcal{L}_f$ is the mean absolute error between the predicted and ground-truth future endogenous values. The intrinsic and cross-type streams attend to different token groups, but their outputs can still be aligned. We therefore penalize the absolute cosine similarity between their pooled outputs to discourage this alignment.

$$\mathcal{L} = \mathcal{L}_f + \lambda_{\text{orth}} \mathcal{L}_{\text{orth}}, \quad \mathcal{L}_{\text{orth}} = \frac{1}{L} \sum_{l=1}^{L} \mathbb{E}\Big[\Big|\cos\big(\mathbf{o}^{\text{int},(l)}, \, \mathbf{o}^{\text{cross},(l)}\big)\Big|\Big], \tag{11}$$

where $\mathbf{o}^{\text{int},(l)}, \mathbf{o}^{\text{cross},(l)}$ are the pooled stream outputs of block l as in the gate of Section 2.5, and the expectation is taken over the samples in the mini-batch.

3 Experiments

Datasets. We evaluate on the 12 real-world TSF-X datasets used by DAG (Qiu et al., 2026): NP, PJM, BE, FR, DE, Energy, Sdwpfm1/2, Sdwpfh1/2, Colbun and Rapel. Future exogenous inputs include published forecasts in the five electricity markets, recorded generation in Energy, and re-analysis weather in the four wind-power datasets. See Appendix A.1 for provenance and statistics.

Baselines. We compare MixRoute against 10 baselines that fall into two groups of five. DAG (Qiu et al., 2026), GCGNet (Li et al., 2026), TimeXer (Wang et al., 2024), TFT (Lim et al., 2021) and TiDE (Das et al., 2023) take future exogenous inputs directly. DUET (Qiu et al., 2025), CrossLinear (Zhou et al., 2025a), Amplifier (Fei et al., 2025), TimeKAN (Huang et al., 2025) and PatchTST (Nie et al., 2023) do not natively take future exogenous values. The latter five accept future exogenous variables through the MLP fusion adaptation described in Appendix A.2. The baseline results in Tables 1 and 2 are taken directly from DAG (Qiu et al., 2026); we do not rerun the baselines for these comparisons. At prediction time, all ten baselines have the same input-variable

Table 2: Results on 12 real-world TSF-X datasets without future exogenous variables, where the inputs are only ($\mathbf{X}_{\text{endo}}$, $\mathbf{X}_{\text{exog}}$). All values are averaged over the two forecasting horizons. Red: the best, Blue: the 2nd best. Full per-horizon results: Table 6 in Appendix C.3.

Models	**MixRoute**		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	**0.401**	**0.367**	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	**0.125**	0.211	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	**0.452**	**0.297**	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	**0.430**	**0.244**	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	**0.596**	**0.484**	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	**0.147**	**0.296**	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	0.710	0.587	**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	**0.776**	**0.641**	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	**0.727**	0.634	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	0.876	0.703	**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	**0.110**	0.169	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	**0.232**	**0.269**	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	**10**	**7**	2	4	0	1	0	0	0	0	0	0	0	2	0	0	0	0	1	1	0	0

availability as MixRoute: $\mathbf{X}_{\text{endo}}$, $\mathbf{X}_{\text{exog}}$ and $\mathbf{Y}_{\text{exog}}$ in the main setting of Table 1, and $\mathbf{X}_{\text{endo}}$ and $\mathbf{X}_{\text{exog}}$ alone in the without-future setting of Section 3.2.2.

Setup. We use Mean Squared Error (MSE) and Mean Absolute Error (MAE) on the normalized series as metrics. Every method is evaluated in one short-term and one long-term setting. We evaluate MixRoute using the TFB benchmark framework (Qiu et al., 2024), which also underlies the baseline results reported by DAG (Qiu et al., 2026). We follow the benchmark settings reported by DAG for the datasets, chronological splits, look-back windows, forecasting horizons, and dataset-level standardization; details are provided in Appendix A.3. Endogenous patches are non-overlapping, so MixRoute effectively observes $T_e = P\lfloor T/P \rfloor$ historical steps, never more than the look-back window given to the baselines.

3.1 Main Results

Table 1 reports MSE and MAE for every dataset, averaged over the two forecasting horizons. MixRoute ranks first on all 12 datasets on MSE and on 10 of 12 on MAE. Compared with the strongest baseline DAG, MixRoute reduces MSE by 8.0% and MAE by 4.1% on average; both are means over the twelve per-dataset reductions. Against the joint-modeling baseline GCGNet, the MSE and MAE reductions are 11.9% and 7.4%, respectively. On the four wind-power datasets, the MSE reductions relative to DAG and GCGNet do not exceed 5.2%. Relative to DAG, the MSE reduction exceeds 10% on NP, FR, Energy, Colbun and Rapel, with the largest reduction of 20.2% on NP. Relative to GCGNet, the MSE reduction exceeds 10% on these five datasets and DE, with the largest reduction of 35.3% on Rapel. Appendix C.1 plots the forecast curves of MixRoute, DAG, GCGNet and TimeXer against the ground truth.

3.2 Model Analyses

3.2.1 Ablation Studies

We ablate the variants in Table 3, grouped by the component they modify: token embedding (TE), Mixed Parameterization (MP) and Adaptive Information Routing (AIR). We compare the global-token prediction head with two variants that instead read endo-token states. Removing the global tokens increases the average MSE by 23.7%, and replacing only the head increases it by 22.6%. These are the two largest increases in Table 3. This comparison favors using global tokens rather than endo-token states as the input to the prediction head. The remaining variants retain the global-token prediction head. The full model has the lowest MSE averaged over the four datasets.

1) Removing only the type embedding increases the average MSE. This suggests that type embeddings provide useful information in addition to the type-specific parameter sharing and attention structure. 2) Sharing either the FFN or the attention projections across token types degrades performance, and sharing both increases the average MSE by 8.1%. Across the eight tasks in Table 3, Shared FFN removes more parameters than Shared Attention on average, yet the two variants have similar average MSE. 3) All seven AIR variants have a higher average MSE than the full model in

Table 3: Ablation studies for MixRoute. All values are averaged over the two forecasting horizons of each dataset, and the **Avg** column averages over the four datasets. The best result in each column is in bold. Per-horizon results on all 12 datasets are in Table 8 of Appendix C.5.

Type	Variant	NP		PJM		DE		Energy		Avg	
		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
TE	w/o Type Embedding	0.297	0.310	0.093	0.187	0.364	0.367	0.123	0.266	0.219	0.283
	w/o Global Tokens	0.341	0.337	0.101	0.195	0.414	0.400	0.177	0.323	0.258	0.314
	Global Tokens not in Head	0.346	0.335	0.095	0.190	0.404	0.394	0.179	0.323	0.256	0.310
MP	Shared FFN	0.303	0.311	0.092	0.186	0.357	0.367	0.126	0.275	0.219	0.285
	Shared Attention	0.303	0.315	0.096	0.188	0.355	0.366	0.116	0.259	0.218	0.282
	Fully Shared	0.316	0.321	0.097	0.189	0.366	0.373	0.123	0.269	0.226	0.288
AIR	w/o AIR	0.296	0.310	0.094	0.187	0.368	0.372	0.114	0.257	0.218	0.282
	w/o Intrinsic Stream	0.312	0.319	0.098	0.189	0.369	0.374	**0.109**	**0.252**	0.222	0.283
	w/o Cross-Type Stream	0.299	0.309	0.094	0.186	0.362	0.372	0.116	0.260	0.218	0.282
	w/o Gate	0.299	0.314	0.098	0.190	0.366	0.371	0.117	0.262	0.220	0.284
	w/ Constant Gate	0.299	0.311	0.097	0.189	0.354	0.363	0.116	0.261	0.217	0.281
	w/o Causal Mask	0.305	0.315	0.095	0.186	0.356	0.365	0.129	0.279	0.221	0.286
	w/o Orthogonality Loss	0.295	0.310	0.095	0.187	0.360	0.369	0.117	0.263	0.216	0.282
	w/o MP and AIR	0.308	0.314	0.095	0.187	0.366	0.374	0.127	0.272	0.224	0.287
	MixRoute	**0.289**	**0.308**	**0.090**	**0.184**	**0.347**	**0.362**	**0.109**	**0.252**	**0.209**	**0.276**

Table 3. Removing the gate sets $\mathbf{g} = 1$, so the two stream outputs are added without learned weighting. The higher average MSE indicates that learned gating improves over unweighted summation in this comparison. We also replace the input-dependent gate with a learned constant for each head in each layer. This variant has higher average MSE on all four datasets, supporting input-dependent gating over learned constant scaling. Removing the cross-type stream only mildly increases the average MSE. Global tokens still attend to exog-tokens in every block and feed the prediction head. Because this information path remains available, the small increase does not imply that exogenous information is unimportant. Allowing the intrinsic stream to attend to all endogenous patches without a causal mask increases the average MSE. This suggests that the causal mask provides a benefit beyond the order information encoded by RoPE. Removing the orthogonality penalty while leaving the two-stream structure intact degrades performance. 4) Removing both Mixed Parameterization and AIR increases the average MSE by 7.4%, even with the global-token prediction head retained. Thus, the head alone does not account for the full model's performance.

3.2.2 Forecasting without Future Exogenous Variables

When $\mathbf{Y}_{\text{exog}}$ is withheld from every method, MixRoute ranks first on 10 of the 12 datasets in MSE and 7 in MAE (Table 2). MixRoute reduces MSE by 2.3% on average relative to DAG and by 4.0% relative to GCGNet. In this setting no component of MixRoute is added, removed or replaced. The exogenous projection takes only the historical segment as input (Section 2.3), and the training objective is unchanged. Exog-tokens are indexed by variable rather than by time step, so the cross-type stream attends over the same $D + N_g$ tokens in either setting. Each exog-token now encodes only the historical segment. In contrast, DAG predicts future exogenous values using its Temporal Correlation Module, while GCGNet generates them using its Variational Generator. Both reuse a module from their with-future architectures to estimate future exogenous values, so the endogenous forecast is conditioned on these estimates rather than on observed future exogenous values.

3.2.3 Practical Situations of Exogenous Variables

The exogenous inputs in Table 1 are complete, have the same historical length and sampling rate as the target series, and are aligned with it at each time step. In practice, exogenous data can violate any of these conditions. Heterogeneity therefore also arises from data availability, sampling and alignment. Following TimeXer (Wang et al., 2024), we test four practical settings on DE, Energy, NP and PJM, each violating one of the above conditions. We follow the protocol of Table 1; look-back lengths are nominal, before patch truncation. Figure 3 reports the MSE averaged over the four datasets, and Appendix C.4 reports each dataset separately.

Missing exogenous values. Sensor faults leave gaps in the exogenous record. We mask a fraction $\rho \in \{10\%, 30\%, 50\%\}$ of the exogenous values at random, over both the historical and the future segments. Masked values are imputed with the channel's training mean (Zero), or with that mean plus unit-standard-deviation noise (Random). Every model is retrained under the same corruption.

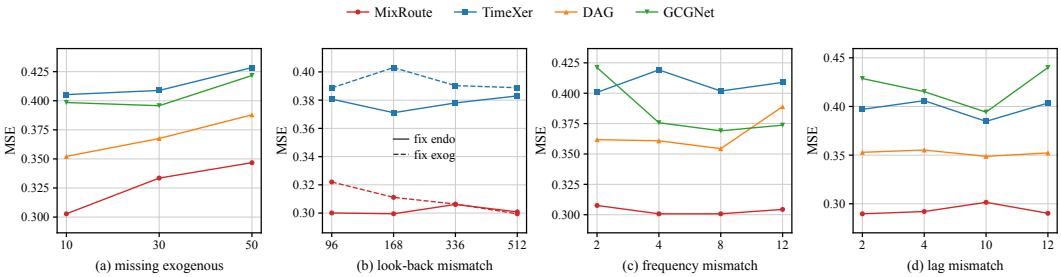


Figure 3: Forecasting performance (MSE) under missing exogenous values in (a), mismatched look-back lengths in (b), mismatched sampling rates in (c) and misaligned exogenous variables in (d), all at $F = 360$. In (a), MSE is averaged across the two imputation methods.

All models receive the identical mask within a condition, at both training and inference. MixRoute has the lowest average MSE at every masking ratio in Figure 3(a), and this average increases with ρ.

Mismatched look-back lengths. The endogenous and the exogenous series do not always have the same amount of available history. MixRoute represents each exogenous variable separately, so the two look-back lengths need not be equal. We run two settings. The first uses a fixed endogenous look-back window of 720 steps and truncates the exogenous look-back window to $\{96, 168, 336, 512\}$ steps. The second uses a fixed exogenous look-back window of 720 steps and truncates the endogenous one to the same values. DAG and GCGNet require equal endogenous and exogenous look-back lengths, so both settings apply only to MixRoute and TimeXer. MixRoute achieves lower average MSE than TimeXer at every truncated look-back length in both settings (Figure 3(b)).

Mismatched sampling rates. Exogenous variables are often recorded at a lower sampling rate than the target series. For $k \in \{2, 4, 8, 12\}$ we keep every k-th step of every exogenous variable. We apply this operation to both the historical and the future segments. The target series remains at hourly resolution. MixRoute and TimeXer read the resulting shorter series directly. DAG and GCGNet require step-aligned input. For these two models, we fill the dropped time steps with the value at the nearest retained time step (Wang et al., 2024). MixRoute achieves the lowest average MSE at every downsampling factor k in Figure 3(c). MixRoute's MSE, averaged over the four datasets, rises by at most 7% relative to the original sampling rate.

Misaligned exogenous variables. Exogenous readings can arrive later than the endogenous series. For $\delta \in \{2, 4, 10, 12\}$, we delay every exogenous variable by δ time steps. The historical and the future segments are shifted together. The first δ time steps are filled with the first observed value. The endogenous series and the timestamps remain unchanged. All models receive the same shifted exogenous record. We also assign each exogenous variable a deterministic delay from $\{2, 4, 8, 12\}$ according to the order of the variables. No single global realignment can correct this per-variable setting. MixRoute achieves the lowest average MSE at every uniform delay in Figure 3(d). MixRoute's MSE, averaged over the four datasets, stays within 5% of the aligned setting.

4 Conclusion

We propose MixRoute, a unified Transformer framework for TSF-X that addresses two design issues in joint modeling: homogeneous parameterization and entangled attention. Mixed Parameterization shares parameters across endogenous patches and gives each exogenous and global token a private set, while Adaptive Information Routing forks the endogenous query into an intrinsic and a cross-type stream and fuses them with a learned gate. Extensive experiments on 12 real-world TSF-X datasets show that MixRoute ranks first overall in both the with-future and without-future settings. When future exogenous values are unavailable, the margin narrows, yet MixRoute needs no architectural change. Future work includes characterizing when exogenous variables reward type-aware routing, extending MixRoute to static, categorical and irregularly sampled exogenous inputs, and probabilistic forecasting.

AI USE STATEMENT

We used large language models to assist with code debugging, language and formatting checks, and pre-submission consistency checks. The methodology, experimental analysis, conceptual contributions and conclusions were all developed by the authors. We reviewed the models' output and take full responsibility for this manuscript.

ETHICS STATEMENT

This study relies solely on publicly available datasets that contain no personally identifiable information. The proposed framework is designed for socially beneficial applications in time series forecasting. No human subjects were involved in this research.

REPRODUCIBILITY STATEMENT

All experimental results in this work can be reproduced. Our model code is released in an anonymous repository:
`https://anonymous.4open.science/r/MixRoute-release-4EBB`.

REFERENCES

Wuzhida Bao, Yuting Cao, Yin Yang, Hangjun Che, Junjian Huang, and Shiping Wen. Data-driven stock forecasting models based on neural networks: A review. *Information Fusion*, 113:102616, 2025.

Konstantinos Benidis, Syama Sundar Rangapuram, Valentin Flunkert, Yuyang Wang, Danielle Maddix, Caner Turkmen, Jan Gasthaus, Michael Bohlke-Schneider, David Salinas, Lorenzo Stella, et al. Deep learning for time series forecasting: Tutorial and literature survey. *ACM Computing Surveys*, 55(6):1–36, 2022.

Xinyang Chen, Huidong Jin, Yu Huang, and Zaiwen Feng. Xlinear: a lightweight and accurate mlp-based model for long-term time series forecasting with exogenous inputs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 20325–20335, 2026.

Yipeng Chen, Heng Zhang, Xingyou Yan, and Qiang Miao. Supply chain demand forecasting based on multi-time scale data fusion network. *Computers & Industrial Engineering*, 207:111324, 2025.

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. ISSN 2835-8856.

Qi Dong, Rubing Huang, Chenhui Cui, Dave Towey, Ling Zhou, Jinyu Tian, and Jianzhou Wang. Short-term electricity-load forecasting by deep learning: A comprehensive survey. *Engineering Applications of Artificial Intelligence*, 154:110980, 2025.

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

Huan Gui, Ruoxi Wang, Ke Yin, Long Jin, Maciej Kula, Taibai Xu, Lichan Hong, and Ed H Chi. Hiformer: Heterogeneous feature interactions learning with transformers for recommender systems. *arXiv preprint arXiv:2311.05884*, 2023.

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.

Songtao Huang, Zhen Zhao, Can Li, and Lei Bai. Timekan: Kan-based frequency decomposition learning architecture for long-term time series forecasting. In *International Conference on Learning Representations*, volume 2025, pp. 28514–28529, 2025.

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.

Yu-Hsiang Lan and Eric K Oermann. Gateformer: Advancing multivariate time series forecasting through temporal and variate-wise attention with gated representations. *arXiv preprint arXiv:2505.00307*, 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 *International Conference on Learning Representations*, volume 2026, pp. 27904–27926, 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.

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*, volume 2024, pp. 11116–11140, 2024.

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

Yuhang Niu, Jiaqi Ye, Shubao Zhao, Zhaoxiang Hou, Chengyi Yang, Zengxiang Li, Yanlong Wen, and Xiaojie Yuan. Hclts: Mining customers' consumption patterns in natural gas time series with hierarchical contrastive learning. In *ICASSP 2025-2025 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pp. 1–5. IEEE, 2025.

Dalin Qin, Guobing Liu, Zengxiang Li, Weicheng Guan, Shubao Zhao, and Yi Wang. Federated deep contrastive learning for mid-term natural gas demand forecasting. *Applied Energy*, 347: 121503, 2023.

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, May 2024. ISSN 2150-8097.

Xiangfei Qiu, Xingjian Wu, Yan Lin, Chenjuan Guo, Jilin Hu, and Bin Yang. Duet: Dual clustering enhanced multivariate time series forecasting. In *Proceedings of the 31st ACM SIGKDD conference on knowledge discovery and data mining v. 1*, pp. 1185–1196, 2025.

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 *Forty-third International Conference on Machine Learning*, 2026.

Helder Relvas. Daily air pollution, meteorology, and hospitalization dataset – Lisbon, 2026. URL `https://zenodo.org/doi/10.5281/zenodo.19339816`. Creative Commons Attribution 4.0 International.

Lei Ren, Zidi Jia, Yuanjun Laili, and Di Huang. Deep learning for time-series prediction in iiot: progress, challenges, and prospects. *IEEE transactions on neural networks and learning systems*, 35(11):15072–15091, 2023.

Minlan Shao, Zijian Zhang, Yili Wang, Yiwei Dai, Xu Shen, and Xin Wang. Hyperd: Hybrid periodicity decoupling framework for traffic forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 15689–15697, 2026.

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 *International conference on learning representations*, volume 2025, pp. 34635–34667, 2025.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

Peiwang Tang and Weitai Zhang. Unlocking the power of patch: Patch-based mlp for long-term time series forecasting. In *Proceedings of the AAAI conference on artificial intelligence*, volume 39, pp. 12640–12648, 2025.

Kshitij Tayal, Arvind Renganathan, Xiaowei Jia, Vipin Kumar, and Dan Lu. Exotst: exogenous-aware temporal sequence transformer for time series prediction. In *2024 IEEE International Conference on Data Mining (ICDM)*, pp. 857–862. IEEE, 2024.

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. *Advances in neural information processing systems*, 37:469–498, 2024.

Haixu Wu, Jiehui Xu, Jianmin Wang, and Mingsheng Long. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. *Advances in neural information processing systems*, 34:22419–22430, 2021.

Guoqi Yu, Juncheng Wang, Chen Yang, Jing Qin, Angelica Aviles-Rivero, and Shujun Wang. Decentralized attention fails centralized signals: Rethinking transformers for medical time series. In *International Conference on Learning Representations*, volume 2026, pp. 108935–108955, 2026.

Liren Yu, Wenming Zhang, Silu Zhou, Tao Zhang, Zhixuan Zhang, and Dan Ou. Hhft: Hierarchical heterogeneous feature transformer for recommendation systems. *arXiv preprint arXiv:2511.20235*, 2025.

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

Ping Zhang, Jiayu Leng, Liang Yang, Anchen Li, Xiaochun Cao, and Riting Xia. Multi-source information driven spatio-temporal hypergraph learning for traffic forecasting. In *Proceedings of the ACM Web Conference 2026*, pp. 580–591, 2026a.

Zhaoqi Zhang, Haolei Pei, Jun Guo, Tianyu Wang, Yufei Feng, Hui Sun, Shaowei Liu, and Aixin Sun. Onetrans: Unified feature interaction and sequence modeling with one transformer in industrial recommender. In *Proceedings of the ACM Web Conference 2026*, pp. 8162–8170, 2026b.

Shubao Zhao, Ming Jin, Zhaoxiang Hou, Chengyi Yang, Zengxiang Li, Qingsong Wen, and Yi Wang. Himtm: hierarchical multi-scale masked time series modeling with self-distillation for long-term forecasting. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, pp. 3352–3362, 2024a.

Shubao Zhao, Xinxing Zhou, Ming Jin, Zhaoxiang Hou, Chengyi Yang, Zengxiang Li, Qingsong Wen, Yi Wang, Yanlong Wen, and Xiaojie Yuan. Rethinking self-supervised learning for time series forecasting: A temporal perspective. *Knowledge-Based Systems*, 305:112652, 2024b.

Pengfei Zhou, Yunlong Liu, Junli Liang, Qi Song, and Xiangyang Li. Crosslinear: Plug-and-play cross-correlation embedding for time series forecasting with exogenous variables. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 2*, pp. 4120–4131, 2025a.

Xinxing Zhou, Jiaqi Ye, Shubao Zhao, Ming Jin, Zhaoxiang Hou, Chengyi Yang, Zengxiang Li, Yanlong Wen, and Xiaojie Yuan. Noise-robust and sector-aware representation learning for natural gas demand forecasting. *Expert Systems with Applications*, pp. 129964, 2025b.

Table 4: Statistics of the 12 datasets. #Ex. is the number of exogenous variables.

Dataset	#Ex.	Exogenous variables	Endogenous variable	Frequency	Length	Split
NP	2	grid load, wind power forecasts	Nord Pool electricity price	1 hour	52,416	7:1:2
PJM	2	system load, COMED zonal load forecasts	COMED zonal electricity price	1 hour	52,416	7:1:2
BE	2	generation, system load forecasts	Belgian electricity price	1 hour	52,416	7:1:2
FR	2	generation, system load forecasts	French electricity price	1 hour	52,416	7:1:2
DE	2	aggregated wind and solar generation, Amprion zonal load forecasts	German electricity price	1 hour	52,416	7:1:2
Energy	5	battery, geothermal, hydroelectric, solar, wind generation	thermoelectric generation	1 hour	13,064	7:1:2
Sdwpfm1/2	6	temperature, surface pressure, relative humidity, wind speed, wind direction, total precipitation	wind turbine active power	30 minutes	29,281	7:1:2
Sdwpfh1/2	6	temperature, surface pressure, relative humidity, wind speed, wind direction, total precipitation	wind turbine active power	1 hour	14,641	7:1:2
Colbun	2	precipitation, tributary inflow	reservoir water level	1 day	2,958	7:1:2
Rapel	2	precipitation, tributary inflow	reservoir water level	1 day	3,366	7:1:2

A Experimental Details

A.1 Datasets

The 12 datasets comprise five from a public electricity price forecasting benchmark (Lago et al., 2021) (NP, PJM, BE, FR, DE) and seven collected by DAG (Qiu et al., 2026) (Energy, Sdwpfm1, Sdwpfm2, Sdwpfh1, Sdwpfh2, Colbun, Rapel). NP records the Nord Pool market, with the electricity price as the endogenous series and grid load and wind power forecasts as the exogenous variables. PJM records the Commonwealth Edison (COMED) zone of the Pennsylvania-New Jersey-Maryland interconnection, with the zonal electricity price as the endogenous series and system load and COMED zonal load forecasts as the exogenous variables. BE records the Belgian market, with the electricity price as the endogenous series and forecasts of generation and system load in the neighboring French system as the exogenous variables. FR records the French market, with the electricity price as the endogenous series and generation and system load forecasts as the exogenous variables. DE records the German market, with the electricity price as the endogenous series and forecasts of aggregated wind and solar generation and Amprion zonal load as the exogenous variables. In these five markets, the exogenous series are published forecasts (Wang et al., 2024) rather than measurements, so their future values are available at prediction time.

Energy records power generation in the Chilean national grid, with thermoelectric output as the endogenous series and battery, geothermal, hydroelectric, solar and wind generation as the exogenous variables (Qiu et al., 2026). The future exogenous values here are recorded generation rather than published forecasts. Sdwpfm1, Sdwpfm2, Sdwpfh1 and Sdwpfh2 record the active power of two turbines of one wind farm, with temperature, surface pressure, relative humidity, wind speed, wind direction and total precipitation from the ERA5 reanalysis (Hersbach et al., 2020) as the exogenous variables. Both turbines use the same exogenous variables, and the supplied future values are weather reanalysis data rather than weather forecasts. Colbun and Rapel are two Chilean hydropower reservoirs, with water level as the endogenous series and precipitation and tributary inflow as the exogenous variables (Qiu et al., 2026).

A.2 Baselines

The baseline results in Tables 1 and 2 are taken directly from DAG (Qiu et al., 2026). These comparisons evaluate forecasting performance against published results, rather than under matched hyperparameter-search budgets. The 10 baselines fall into two groups according to whether they natively support future exogenous values.

Native support for future exogenous values. DAG (Qiu et al., 2026) discovers temporal and cross-variable correlations in two parallel encoders and combines their predictions by a fixed weighted sum. GCGNet (Li et al., 2026) embeds endogenous and exogenous variables in one token space and constrains their interaction with a graph-consistency objective. TimeXer (Wang et al., 2024) patches the endogenous series, embeds each exogenous variable as a single token, and links the two through one endogenous global token that queries the exogenous tokens by cross-attention. TFT (Lim et al., 2021) filters static and time-varying covariates with variable selection networks and attends over the forecasting horizon. TiDE (Das et al., 2023) is an MLP encoder-decoder that incorporates covariates by concatenating them with the encoded history at each step.

No native support for future exogenous values. PatchTST (Nie et al., 2023) segments the series into patches and applies a channel-independent Transformer. DUET (Qiu et al., 2025) clusters series

both by temporal pattern and by channel and assigns a separate expert to each cluster. CrossLinear (Zhou et al., 2025a) equips a linear predictor with a plug-and-play cross-correlation embedding. Amplifier (Fei et al., 2025) amplifies the low-energy frequency components that a plain squared-error objective tends to neglect. TimeKAN (Huang et al., 2025) decomposes the series by frequency and learns each band with Kolmogorov-Arnold layers.

Adaptation of the second group. In the evaluation protocol of DAG (Qiu et al., 2026), these five models incorporate future exogenous values through MLP fusion. The backbone $\mathcal{F}_{\text{base}}$ is first applied to the historical inputs alone. At each forecast step, its output is concatenated with the corresponding future exogenous values and passed through a two-layer MLP to produce the endogenous forecast:

$$\hat{\mathbf{Y}}_{\text{endo}} = \text{MLP}(\text{Concat}(\mathcal{F}_{\text{base}}(\mathbf{X}_{\text{endo}}, \mathbf{X}_{\text{exog}}),\ \mathbf{Y}_{\text{exog}}))\,. \tag{12}$$

The five models in the first group require no such adaptation, as they take $\mathbf{Y}_{\text{exog}}$ as input directly. In the without-future setting (Section 3.2.2), future exogenous values are unavailable at prediction time, and the MLP fusion is not used.

A.3 Implementation Details

Evaluation protocol. For the 12 TSF-X datasets, we partition each time series chronologically into training, validation, and test segments in a 7:1:2 ratio, following DAG (Qiu et al., 2026). Each variable is standardized using statistics estimated only from the training segment, and the resulting transformation is applied to the validation and test segments. We report MSE and MAE for the endogenous target in the normalized space. Evaluation uses rolling forecasts with a stride of one time step. All test batches, including the final incomplete batch, are retained, consistent with the no-"Drop Last" setting reported by DAG.

Forecast settings. Every method is evaluated in one short-term and one long-term setting per dataset. On the 10 hourly and half-hourly datasets, the short-term setting forecasts 24 steps from a look-back window of 168 steps, and the long-term setting forecasts 360 steps from a look-back window of 720 steps. For the daily Colbun and Rapel datasets, the short-term setting forecasts 10 steps from a look-back window of 60 steps, and the long-term setting forecasts 30 steps from a look-back window of 180 steps.

Training. MixRoute is implemented in PyTorch and trained with AdamW on NVIDIA RTX 4090 GPUs. The forecasting loss is L1. We use validation loss for early stopping and checkpoint selection.

B Related Work

B.1 Time Series Forecasting

Deep forecasting models use recurrent, convolutional, Transformer and MLP architectures (Benidis et al., 2022). Transformer-based methods differ mainly in tokenization and attention design. Autoformer (Wu et al., 2021) combines series decomposition with an auto-correlation mechanism that replaces standard self-attention. PatchTST (Nie et al., 2023) segments the series into patches and applies channel-independent Transformers. HiMTM (Zhao et al., 2024a) models the series at multiple scales and pre-trains the representation by masked reconstruction with self-distillation. iTransformer (Liu et al., 2024) inverts the Transformer architecture by treating each variable as a token to capture multivariate correlations. DUET (Qiu et al., 2025) clusters series both by temporal pattern and by channel and assigns a separate expert to each cluster. Gateformer (Lan & Oermann, 2025) first gates patch-attention and whole-series embeddings to form per-variate representations. It then applies variate-wise attention and uses a second gate to combine channel-independent and channel-interacting representations.

MLP-based methods offer alternatives to attention-based forecasting. DLinear (Zeng et al., 2023) demonstrates that a simple linear model can outperform many Transformer variants by decomposing series into trend and seasonal components. PatchMLP (Tang & Zhang, 2025) argues that the patching mechanism, rather than self-attention, is the key driver of forecasting improvements. Amplifier (Fei et al., 2025) amplifies the low-energy frequency components that a plain squared-error objective tends to neglect. Beyond MLPs, TimeKAN (Huang et al., 2025) decomposes the series

by frequency and learns each band with Kolmogorov-Arnold layers. Another line of work explicitly models exogenous variables.

B.2 Time Series Forecasting with Exogenous Variables

Deep forecasting models with exogenous variables fall into two families: decoupled modeling and joint modeling. Decoupled modeling methods assign temporal and cross-variable correlations to separate modules. TFT (Lim et al., 2021) introduces variable selection networks and multi-horizon attention to integrate static and temporal covariates. TiDE (Das et al., 2023) employs an MLP-based encoder-decoder that incorporates future covariates through concatenation. CrossLinear (Zhou et al., 2025a) captures cross-variable correlations via plug-and-play cross-correlation embedding. TimeXer (Wang et al., 2024) patches the endogenous series and embeds each exogenous variable as one token. The two token types are then connected through a single endogenous global token, which queries the exogenous tokens through cross-attention. The exogenous tokens are not updated, and the endogenous patch tokens access exogenous information only indirectly through the global token. XLinear (Chen et al., 2026) also routes exogenous information through endogenous global tokens, but replaces attention with MLP-based gating along the temporal and variable dimensions. Exogenous information reaches its prediction head only through these tokens. DAG (Qiu et al., 2026) instead runs the temporal and cross-variable modules in parallel. Its correlation injection modules use input-dependent gates to combine attention scores computed with their own query and key projections with those transferred from correlation discovery. The final forecasts of the temporal and cross-variable modules are then combined by a fixed weighted sum. AIR instead computes separately normalized attention over endogenous tokens and over exogenous and global tokens using distinct endogenous queries, then fuses the stream outputs within each block. Joint modeling approaches remove this separation by embedding all variables in a single token space. This allows temporal and cross-variable correlations to be modeled together from the first layer. GCGNet (Li et al., 2026) embeds endogenous and exogenous variables into a shared token space and models their interactions through graph-consistent constraints.

Type-aware parameterization has precedents in recommender systems. HiFormer (Gui et al., 2023) and OneTrans (Zhang et al., 2026b) assign parameters by token type, whereas Time-MoE (Shi et al., 2025) uses mixture-of-experts to learn parameter assignment from token content in time series forecasting. HHFT (Yu et al., 2025) assigns parameters per semantic feature block. MixRoute combines Mixed Parameterization with AIR within each attention block. AIR uses separate endogenous queries for temporal modeling and exogenous information acquisition. Unlike the routing in Time-MoE, the gate in AIR does not select which modules to activate. Both streams remain active. Gateformer's gates combine representations from staged temporal and variate-wise modeling, whereas AIR gates the cross-type attention output before adding it to the intrinsic output within each block.

C Experimental Results

C.1 Visualization of Prediction Results

Figure 4 compares MixRoute with DAG, the strongest baseline in Table 1, as well as GCGNet and TimeXer. For both NP and Energy, we show one test window at each forecasting horizon, $F = 24$ and $F = 360$. All four models recover the periodic shape of the target, but their forecasts differ in mean level. MixRoute generally follows the ground truth more closely, with smaller overall offsets. On NP at $F = 360$, for example, GCGNet captures the recurring fluctuations but shows a more noticeable overall offset than MixRoute. On Energy at $F = 360$, DAG also follows the repeated rises and falls, yet its predictions are less closely aligned with the ground truth than those of MixRoute.

C.2 Gate Visualization

Figure 5 shows the AIR gate $\mathbf{g}$ defined in Section 2.5 on NP, PJM, DE and Energy at $F = 24$ in the with-future setting. For each dataset, we plot all eight heads in the final encoder layer over the first 200 chronological test windows. Within each dataset, the gate assigns different weights to the cross-type stream across heads. On NP, PJM and DE, the gate values of several heads also vary across

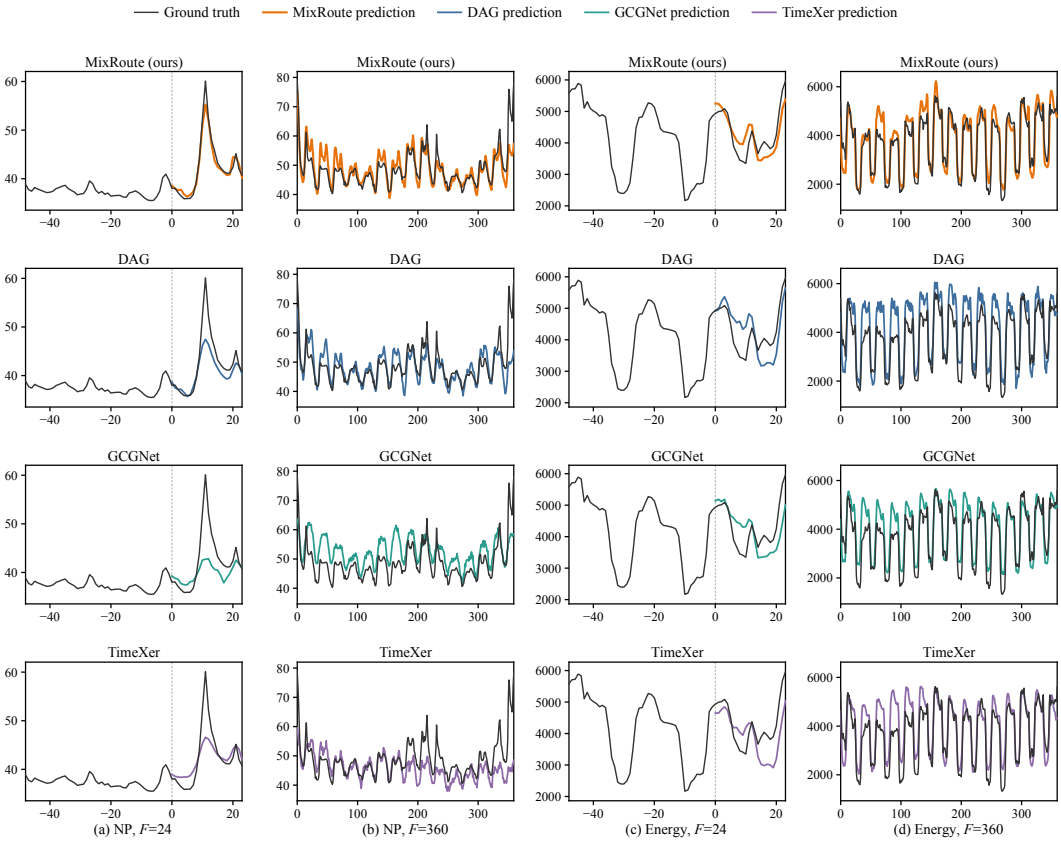


Figure 4: Forecasts of MixRoute, DAG, GCGNet and TimeXer on NP and Energy, compared with the ground truth. Each row corresponds to one model, and each column shows one test window for a dataset and forecasting horizon. For $F = 24$, the 48 historical time steps preceding the forecast are shown to the left of the dashed line. For $F = 360$, only the forecast interval is shown.

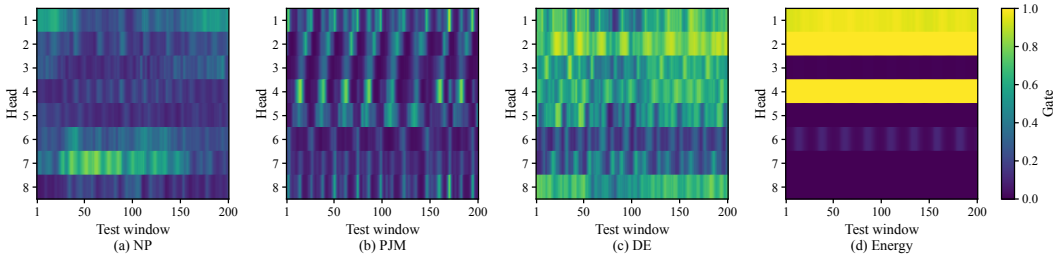


Figure 5: AIR gate values in the final encoder layer on (a) NP, (b) PJM, (c) DE and (d) Energy at $F = 24$ in the with-future setting. Rows correspond to the eight attention heads, and columns to the first 200 chronological test windows. All four heatmaps share a $[0, 1]$ colour scale.

windows. On Energy, most heads remain near zero or one throughout the displayed windows, with little variation.

C.3 Full Results

Tables 5 and 6 present the full forecasting results with and without future exogenous variables, respectively. For each dataset, the tables report short- and long-term forecasting results, followed by an Avg row showing the average across the two horizons. These averages are reported in Tables 1 and 2

Table 5: Full results on 12 real-world TSF-X datasets, where the inputs are ($\mathbf{X}_{\text{endo}}$, $\mathbf{X}_{\text{exog}}$ and $\mathbf{Y}_{\text{exog}}$). The best and second-best results are shown in red and blue, respectively. The 1st Count row aggregates over all 36 rows, crediting every model that attains the best value in a row.

Models		**MixRoute**		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	**0.166**	**0.215**	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	**0.412**	**0.400**	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	**0.289**	**0.308**	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	0.058	0.146	**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	**0.121**	0.222	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	**0.090**	0.184	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	**0.336**	**0.224**	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	**0.462**	**0.307**	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	**0.399**	**0.265**	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	**0.298**	**0.163**	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	**0.440**	**0.256**	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	**0.369**	**0.210**	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	**0.233**	**0.302**	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	**0.461**	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	**0.347**	**0.362**	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	**0.064**	**0.197**	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	**0.155**	**0.308**	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	**0.109**	**0.252**	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	0.372	0.409	**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	**0.461**	0.505	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	**0.416**	**0.457**	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	0.430	0.450	**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	**0.500**	0.527	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	**0.465**	0.489	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	**0.384**	0.444	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	0.480	**0.514**	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	**0.432**	**0.479**	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	0.459	0.489	**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	**0.533**	**0.545**	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	**0.496**	**0.517**	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	**0.048**	**0.083**	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	**0.125**	**0.207**	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	**0.087**	**0.145**	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**	**0.199**	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	**0.249**	**0.321**	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	**0.198**	**0.260**	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		**31**	**25**	4	9	1	3	0	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0

of the main text. The forecasting horizons are 24 and 360 steps for all datasets except Colbun and Rapel. These two datasets are recorded daily and use horizons of 10 and 30 steps (Appendix A.3).

C.4 Practical Situations of Exogenous Variables

This subsection reports the per-dataset results for the four practical situations examined in Section 3.2.3: missing observations and mismatches in look-back length, sampling rate and temporal alignment.

Missing exogenous values. Table 7 reports the forecasting results at each masking ratio as well as the averages across the three ratios, for both $F = 24$ and $F = 360$. For each dataset, the rows are grouped by forecasting horizon. Within each group, the first three rows correspond to masking ratios of 10%, 30% and 50%, and the fourth reports their average.

Mismatched look-back lengths. A newly introduced sensor, for instance, provides only a short exogenous record, so the two look-back lengths can differ in practice. In the two settings described in Section 3.2.3, we fix either the endogenous or the exogenous look-back window at 720 steps and truncate the other. The future exogenous segment remains 360 steps long in both settings. Figure 6 reports the results on DE, Energy, NP and PJM for MixRoute and TimeXer, the only baseline considered here that supports different endogenous and exogenous look-back lengths. MixRoute achieves a lower MSE than TimeXer at every truncated look-back length in both settings on all

Table 6: Full results on 12 real-world TSF-X datasets without future exogenous variables, where the inputs are only ($\mathbf{X}_{\text{endo}}$, $\mathbf{X}_{\text{exog}}$). The best and second-best results are shown in red and blue, respectively. The 1st Count row aggregates over all 36 rows, crediting every model that attains the best value in a row.

Models		**MixRoute**		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	**0.220**	**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	**0.582**	0.479	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	**0.401**	**0.367**	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	**0.070**	0.164	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	**0.179**	**0.258**	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	**0.125**	0.211	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	**0.355**	**0.239**	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	0.549	0.354	**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	**0.452**	**0.297**	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	**0.336**	**0.184**	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	0.524	0.303	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	**0.430**	**0.244**	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**	**0.400**	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	**0.780**	0.568	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	**0.596**	**0.484**	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	0.106	0.251	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	0.188	0.341	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	**0.147**	**0.296**	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	0.540	0.502	**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	0.880	**0.672**	**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	0.710	0.587	**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	**0.609**	0.540	**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	**0.943**	0.743	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	**0.776**	**0.641**	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	**0.635**	**0.574**	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	**0.818**	0.694	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	**0.727**	0.634	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	0.785	**0.633**	**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	0.968	0.772	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	0.876	0.703	**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	**0.060**	**0.085**	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	**0.160**	0.254	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	**0.110**	0.169	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	0.198	0.216	**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	**0.266**	**0.321**	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	**0.232**	**0.269**	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		**25**	**17**	8	8	1	3	2	2	0	0	0	0	0	6	1	2	0	0	2	2	0	0

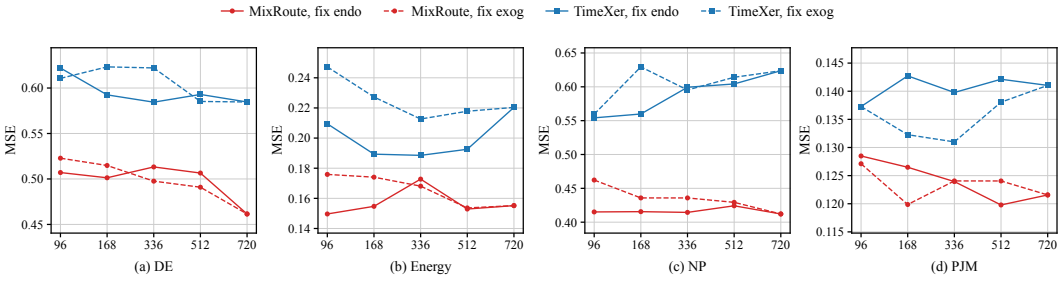


Figure 6: Forecasting performance (MSE) with mismatched look-back lengths and a fixed horizon $F = 360$. Solid lines indicate a fixed endogenous look-back window of 720 steps and varying exogenous look-back lengths. Dashed lines indicate a fixed exogenous look-back window of 720 steps and varying endogenous look-back lengths.

four datasets. These results show that MixRoute can handle endogenous and exogenous series with different look-back lengths.

Table 7: Robustness at horizons $F = 24$ and $F = 360$ with 10%, 30% and 50% of the exogenous values masked at random. The two imputation methods are Zero, which uses the channel's training mean, and Random, which adds unit-standard-deviation noise to that mean. The best of the five models is in bold.

Mask Type			Zero										Random									
Models			MixRoute		DAG		GCGNet		TimeXer		DUET		MixRoute		DAG		GCGNet		TimeXer		DUET	
Metrics			MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
DE	24	10	**0.245**	**0.311**	0.335	0.354	0.303	0.345	0.345	0.365	0.397	0.389	**0.258**	**0.320**	0.325	0.353	0.320	0.359	0.363	0.382	0.414	0.400
		30	**0.275**	**0.325**	0.354	0.366	0.331	0.359	0.373	0.381	0.414	0.399	**0.304**	**0.350**	0.396	0.393	0.380	0.386	0.405	0.407	0.443	0.416
		50	**0.285**	**0.332**	0.377	0.376	0.368	0.375	0.391	0.390	0.439	0.411	**0.345**	**0.376**	0.426	0.414	0.410	0.407	0.480	0.445	0.452	0.419
		Avg	**0.268**	**0.322**	0.355	0.366	0.334	0.360	0.370	0.379	0.417	0.400	**0.303**	**0.349**	0.382	0.387	0.370	0.384	0.416	0.411	0.436	0.412
	360	10	**0.464**	0.427	0.478	**0.417**	0.538	0.458	0.650	0.495	0.645	0.513	0.512	0.447	**0.501**	**0.440**	0.744	0.565	0.652	0.498	0.702	0.537
		30	0.516	0.449	**0.499**	**0.435**	0.561	0.476	0.626	0.484	0.738	0.549	0.576	0.475	**0.551**	**0.464**	0.654	0.521	0.646	0.505	0.791	0.574
		50	0.541	0.460	**0.532**	**0.459**	0.642	0.524	0.686	0.512	0.788	0.570	**0.598**	**0.498**	0.612	**0.498**	0.763	0.568	0.685	0.527	0.858	0.602
		Avg	0.507	0.445	**0.503**	**0.437**	0.580	0.486	0.654	0.497	0.724	0.544	0.562	0.473	**0.554**	**0.467**	0.720	0.551	0.661	0.510	0.784	0.571
Energy	24	10	**0.078**	**0.215**	0.085	0.225	0.089	0.233	0.113	0.263	0.149	0.314	**0.078**	**0.216**	0.083	0.221	0.089	0.233	0.107	0.257	0.143	0.301
		30	**0.077**	**0.213**	0.082	0.221	0.102	0.249	0.110	0.258	0.136	0.296	**0.086**	**0.228**	0.092	0.235	0.092	0.236	0.109	0.259	0.121	0.273
		50	0.096	0.240	**0.087**	**0.227**	0.101	0.246	0.113	0.263	0.130	0.285	**0.096**	0.243	0.098	**0.242**	0.117	0.269	0.123	0.277	0.122	0.274
		Avg	**0.084**	**0.223**	0.085	0.224	0.098	0.243	0.112	0.262	0.138	0.298	**0.087**	**0.229**	0.091	0.233	0.099	0.246	0.113	0.264	0.129	0.283
	360	10	**0.148**	**0.303**	0.206	0.355	0.207	0.362	0.181	0.338	0.302	0.454	**0.167**	**0.322**	0.194	0.342	0.266	0.415	0.191	0.348	0.257	0.408
		30	**0.164**	**0.317**	0.181	0.334	0.239	0.393	0.183	0.339	0.275	0.421	0.202	0.353	**0.192**	**0.342**	0.250	0.399	0.203	0.361	0.217	0.368
		50	**0.165**	**0.319**	0.218	0.365	0.221	0.372	0.185	0.343	0.242	0.390	**0.193**	**0.348**	0.212	0.358	0.237	0.393	0.199	0.360	0.254	0.397
		Avg	**0.159**	**0.313**	0.202	0.351	0.223	0.376	0.183	0.340	0.273	0.422	**0.187**	**0.341**	0.200	0.348	0.251	0.403	0.198	0.356	0.243	0.391
NP	24	10	**0.182**	**0.222**	0.216	0.243	0.231	0.256	0.247	0.272	0.262	0.305	**0.181**	**0.222**	0.222	0.243	0.243	0.261	0.255	0.275	0.259	0.290
		30	**0.188**	**0.228**	0.227	0.248	0.259	0.277	0.258	0.283	0.256	0.289	**0.201**	**0.234**	0.236	0.254	0.264	0.275	0.276	0.288	0.263	0.281
		50	**0.191**	**0.230**	0.236	0.253	0.261	0.275	0.273	0.290	0.260	0.282	**0.214**	**0.244**	0.242	0.260	0.272	0.281	0.304	0.305	0.264	0.278
		Avg	**0.187**	**0.227**	0.226	0.248	0.250	0.269	0.259	0.281	0.260	0.292	**0.198**	**0.234**	0.233	0.252	0.260	0.272	0.278	0.289	0.262	0.283
	360	10	**0.438**	**0.412**	0.575	0.481	0.578	0.481	0.633	0.482	0.616	0.555	**0.442**	**0.413**	0.581	0.483	0.577	0.480	0.652	0.501	0.617	0.546
		30	**0.460**	**0.416**	0.608	0.493	0.583	0.476	0.648	0.497	0.624	0.546	**0.487**	**0.426**	0.621	0.499	0.582	0.480	0.664	0.506	0.619	0.525
		50	**0.487**	**0.424**	0.610	0.495	0.593	0.482	0.683	0.515	0.617	0.525	**0.524**	**0.444**	0.620	0.502	0.605	0.498	0.683	0.515	0.650	0.541
		Avg	**0.461**	**0.417**	0.597	0.490	0.585	0.479	0.655	0.498	0.619	0.542	**0.484**	**0.428**	0.607	0.494	0.588	0.486	0.667	0.507	0.628	0.537
PJM	24	10	0.064	0.152	**0.062**	**0.150**	0.072	0.170	0.077	0.171	0.102	0.215	**0.063**	**0.153**	0.065	0.154	0.077	0.175	0.080	0.175	0.105	0.219
		30	0.068	**0.156**	**0.067**	**0.156**	0.068	0.158	0.085	0.178	0.099	0.206	**0.068**	**0.160**	0.070	0.163	0.070	0.164	0.088	0.183	0.099	0.208
		50	**0.066**	**0.156**	0.071	0.162	0.085	0.185	0.086	0.183	0.100	0.206	0.076	0.171	**0.074**	**0.168**	0.094	0.199	0.094	0.194	0.090	0.193
		Avg	**0.066**	**0.155**	**0.066**	0.156	0.075	0.171	0.083	0.177	0.100	0.209	**0.069**	0.162	0.070	**0.161**	0.080	0.179	0.087	0.184	0.098	0.207
	360	10	**0.126**	0.225	0.141	**0.222**	0.140	0.236	0.140	0.230	0.136	0.235	**0.126**	0.225	0.141	**0.223**	0.138	0.231	0.142	0.232	0.142	0.237
		30	**0.137**	0.232	0.138	**0.223**	0.155	0.248	0.143	0.232	0.146	0.241	**0.128**	**0.226**	0.150	0.232	0.143	0.235	0.157	0.243	0.159	0.252
		50	**0.129**	**0.228**	0.137	0.232	0.145	0.237	0.151	0.240	0.161	0.252	**0.136**	**0.234**	0.161	0.239	0.167	0.254	0.156	0.243	0.171	0.262
		Avg	**0.130**	0.228	0.139	**0.225**	0.147	0.240	0.145	0.234	0.147	0.243	**0.130**	**0.229**	0.151	0.231	0.149	0.240	0.152	0.240	0.157	0.251

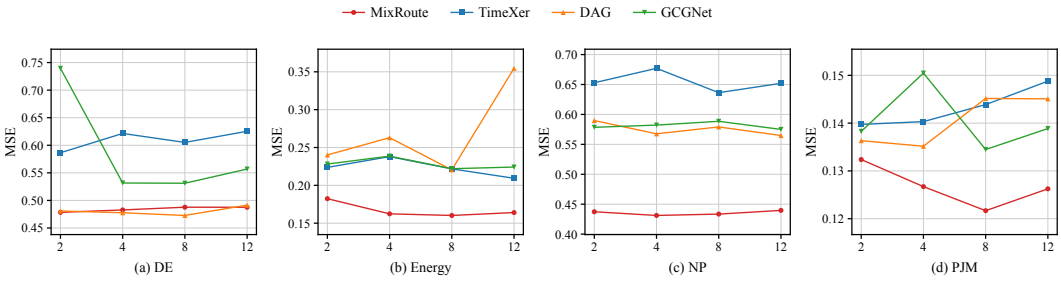


Figure 7: Forecasting performance (MSE) with exogenous variables downsampled by a factor of $k \in \{2, 4, 8, 12\}$ and a fixed horizon $F = 360$. MixRoute and TimeXer use the coarser series directly. For DAG and GCGNet, the downsampled series are upsampled to hourly resolution using the nearest retained value.

Mismatched sampling rates. The four electricity datasets DE, Energy, NP and PJM record both endogenous and exogenous variables hourly, so downsampling the exogenous series simulates a mismatch in sampling rates. Figure 7 reports the results for the four downsampling factors. MixRoute handles all four downsampling factors without upsampling the exogenous series. MixRoute achieves

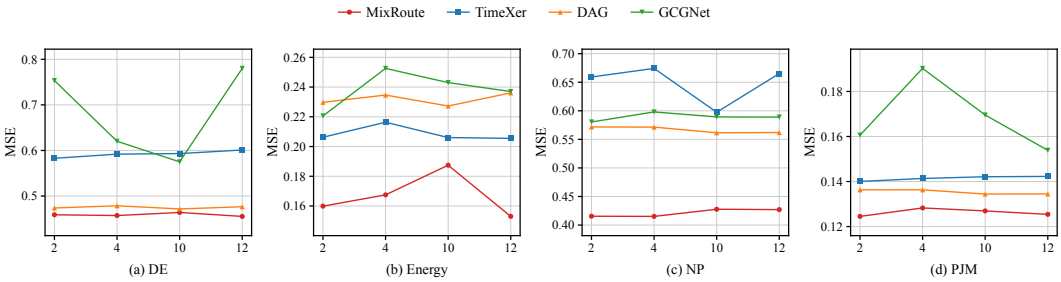


Figure 8: Forecasting performance (MSE) with exogenous variables delayed by δ steps and a fixed horizon $F = 360$.

the lowest MSE in all tested cases except DE at $k = 4$ and $k = 8$. At $k = 12$, MixRoute achieves the lowest MSE on all four datasets.

Misaligned exogenous variables. Figure 8 reports the results for the uniform delays described in Section 3.2.3. MixRoute achieves the lowest MSE on all four datasets at every tested uniform delay and under variable-specific delays. Across the tested uniform delays, MixRoute's MSE changes little on DE, NP and PJM. On Energy, its MSE fluctuates across these delays, but remains lower than that of every baseline.

C.5 Ablation Studies

Table 8 presents the full ablation results on all 12 datasets. For each dataset, results are reported separately for the short-term and long-term forecasting horizons. The four datasets in Table 3 are included in Table 8. The ablation variants and their grouping are the same as in Section 3.2.1. Removing the global tokens increases the average MSE over the 24 tasks by 20.9%, and replacing only the prediction head with one that reads endo-token states increases it by 18.2%. Sharing both the FFN and the attention projections across token types increases the average MSE by 12.2%, while removing Mixed Parameterization and AIR together increases it by 11.6%. The remaining variants increase the average MSE by 2.9% to 8.7%. Across the 24 tasks, MixRoute has the lowest average MSE among all variants. The learnable constant-gate variant has higher MSE than MixRoute on 21 of the 24 tasks. Its input-independent weights, one per head in each layer, use a sigmoid parameterization and are initialized to 0.5.

Table 8: Ablation results on all 12 datasets, reported separately for the two forecasting horizons. F is the forecasting horizon, where **S** denotes the short-term setting (24 steps, or 10 on Colbun and Rapel) and **L** the long-term one (360 steps, or 30 on Colbun and Rapel). For each dataset, forecasting horizon and metric, the best results are shown in bold, including ties.

Type	Variant	F	NP		PJM		BE		FR		DE		Energy		Sdwpfm1		Sdwpfm2		Sdwpfh1		Sdwpfh2		Colbun		Rapel	
			MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
TE	w/o Type Embedding	S	0.176	0.216	0.063	0.151	0.356	0.233	0.309	0.165	0.245	0.307	0.065	**0.197**	0.392	0.417	0.485	0.485	0.424	0.464	0.498	0.512	0.058	0.089	0.171	0.217
		L	0.417	0.405	0.124	0.224	0.489	0.313	0.447	**0.254**	0.484	0.427	0.181	0.335	0.480	0.515	0.582	0.567	0.490	0.529	0.586	0.575	0.144	0.239	0.354	0.381
	w/o Global Tokens	S	0.173	0.218	0.068	0.155	0.363	0.236	0.331	0.173	0.269	0.324	0.105	0.249	0.448	0.456	0.579	0.530	0.501	0.511	0.568	0.543	0.057	0.090	0.187	0.224
		L	0.509	0.455	0.135	0.234	0.489	0.332	0.497	0.291	0.558	0.476	0.248	0.397	0.544	0.532	0.614	0.568	0.571	0.543	0.605	0.574	0.195	0.314	0.324	0.349
	Global Tokens not in Head	S	0.180	0.218	0.065	0.153	0.360	0.234	0.331	0.168	0.267	0.320	0.096	0.239	0.464	0.470	0.538	0.512	0.484	0.504	0.502	0.520	0.064	0.105	0.185	0.224
		L	0.512	0.451	0.126	0.227	0.482	0.331	0.479	0.278	0.541	0.468	0.262	0.406	0.553	0.542	0.593	0.562	0.565	0.550	0.631	0.584	0.221	0.336	**0.242**	0.329
MP	Shared FFN	S	0.170	**0.213**	0.061	0.148	0.348	0.225	0.311	**0.161**	0.238	0.306	0.089	0.234	0.378	0.411	0.439	0.460	0.391	0.447	**0.456**	**0.487**	0.055	0.086	0.215	0.235
		L	0.437	0.408	0.122	0.225	0.483	0.306	0.448	0.259	0.476	0.428	0.164	0.316	0.475	0.517	0.534	0.543	0.486	0.522	0.549	0.550	0.128	0.220	0.311	0.344
	Shared Attention	S	0.168	0.217	0.062	0.150	0.359	0.226	0.314	0.162	0.237	0.303	0.068	0.202	0.391	0.431	0.446	0.463	0.409	0.458	0.469	0.504	0.053	0.085	0.229	0.245
		L	0.438	0.413	0.131	0.227	0.494	0.326	0.487	0.287	0.473	0.429	0.164	0.317	0.494	0.524	0.618	0.590	0.488	0.527	0.568	0.566	0.182	0.266	0.291	0.337
	Fully Shared	S	0.175	0.220	0.069	0.153	0.354	0.225	0.318	0.164	0.241	0.306	0.081	0.220	0.389	0.424	0.464	0.466	0.421	0.461	0.513	0.517	0.055	0.089	0.165	0.210
		L	0.458	0.423	0.126	0.226	0.530	0.341	0.467	0.271	0.491	0.441	0.165	0.319	0.506	0.523	0.607	0.582	0.545	0.558	0.642	0.608	0.135	0.221	0.381	0.388
AIR	w/o AIR	S	0.170	0.215	0.060	0.147	0.343	0.226	0.305	0.163	0.237	0.305	0.068	0.203	0.415	0.423	0.445	0.455	0.390	0.448	0.461	0.495	0.057	0.093	0.171	0.216
		L	0.421	0.405	0.128	0.227	0.471	0.313	0.469	0.270	0.499	0.439	0.159	0.312	0.482	0.516	0.523	0.535	**0.480**	0.518	0.543	0.548	0.121	0.215	0.362	0.383
	w/o Intrinsic Stream	S	0.175	0.215	0.064	0.150	0.337	0.226	**0.296**	0.162	0.239	0.304	0.066	0.199	0.403	0.417	0.447	0.457	0.388	0.448	0.465	0.492	0.056	0.089	0.189	0.216
		L	0.449	0.422	0.131	0.228	0.478	0.320	0.452	0.261	0.499	0.444	**0.151**	**0.304**	0.485	0.514	0.537	0.545	0.482	0.521	0.553	0.555	0.132	0.233	0.371	0.367
	w/o Cross-Type Stream	S	0.170	0.214	0.062	0.147	0.337	0.224	0.299	0.163	0.240	0.305	**0.064**	**0.197**	0.380	0.416	0.447	0.457	0.386	**0.444**	0.493	0.504	0.054	0.088	0.155	0.201
		L	0.428	0.404	0.125	0.224	**0.453**	0.305	0.461	0.269	0.484	0.439	0.169	0.323	0.472	0.513	0.528	0.540	0.482	0.521	0.553	0.555	0.143	0.235	0.303	0.338
	w/o Gate	S	0.172	**0.213**	0.066	0.152	0.333	0.226	0.307	0.162	0.241	0.306	0.068	0.204	0.377	**0.408**	0.482	0.469	0.387	0.448	0.459	0.491	0.050	0.090	0.219	0.247
		L	0.426	0.415	0.131	0.228	0.478	0.304	0.445	0.255	0.492	0.437	0.167	0.321	0.482	0.514	0.518	0.537	0.485	0.524	0.553	0.549	0.128	0.214	0.322	0.348
	w/ Constant Gate	S	0.177	0.217	0.066	0.151	**0.326**	**0.222**	0.304	0.163	0.241	0.305	0.067	0.203	0.378	0.410	0.489	0.472	0.393	0.449	**0.456**	0.490	0.053	0.084	0.158	0.206
		L	0.420	0.404	0.129	0.228	0.476	**0.302**	0.451	0.264	0.467	**0.420**	0.165	0.320	0.472	0.509	0.522	0.537	0.487	0.520	0.553	0.551	**0.113**	0.208	0.354	0.381
	w/o Causal Mask	S	0.173	0.218	0.062	0.147	0.357	0.227	0.315	0.164	0.239	0.305	0.096	0.241	0.387	0.415	0.490	0.485	0.390	0.449	0.461	0.489	0.051	**0.082**	0.174	0.220
		L	0.438	0.413	0.128	0.225	0.470	0.311	0.451	0.263	0.473	0.426	0.162	0.316	0.471	0.510	0.512	0.533	0.488	0.520	0.539	0.548	0.133	0.218	0.283	0.344
	w/o Orthogonality Loss	S	**0.166**	0.215	0.062	0.148	0.333	0.224	0.299	0.163	**0.233**	**0.302**	0.069	0.205	0.377	0.415	0.445	0.456	**0.384**	**0.444**	0.464	0.491	**0.048**	0.083	0.164	0.205
		L	0.423	0.406	0.127	0.227	0.458	0.309	0.466	0.271	0.486	0.436	0.165	0.321	0.474	0.510	0.509	0.536	**0.480**	**0.514**	0.545	0.552	0.132	0.226	0.303	0.349
	w/o MP and AIR	S	0.171	0.215	0.064	0.150	0.353	0.226	0.317	0.163	0.243	0.306	0.085	0.224	0.407	0.423	0.445	0.457	0.440	0.464	0.539	0.522	0.054	0.085	0.183	0.214
		L	0.446	0.413	0.127	0.225	0.501	0.327	0.491	0.287	0.489	0.443	0.169	0.320	0.514	0.532	0.623	0.585	0.526	0.546	0.623	0.601	0.123	0.211	0.322	0.362
	MixRoute	S	**0.166**	0.215	**0.058**	**0.146**	0.336	0.224	0.298	0.163	**0.233**	**0.302**	**0.064**	**0.197**	**0.372**	0.409	**0.430**	**0.450**	**0.384**	**0.444**	0.459	0.489	**0.048**	0.083	**0.147**	**0.199**
		L	**0.412**	**0.400**	**0.121**	**0.222**	0.462	0.307	**0.440**	0.256	**0.461**	0.421	0.155	0.308	**0.461**	**0.505**	**0.500**	**0.527**	**0.480**	**0.514**	**0.533**	**0.545**	0.125	**0.207**	0.249	**0.321**

C.6 Scaling the Fully Shared Variant

The Fully Shared variant shares attention projections and FFNs across token types and has fewer parameters than MixRoute. The higher MSE may therefore be due solely to reduced capacity. To examine this possibility, we increase the capacity of this variant on DE, Energy, NP and PJM. We vary model width, encoder depth or the feed-forward expansion ratio one at a time, up to $4\times$ MixRoute's parameter count.

For every dataset and forecasting horizon shown in Figure 9, at least one scaled configuration achieves lower MSE than the unscaled Fully Shared variant. The gains vary across datasets and scaling directions. On Energy, for example, some deeper configurations achieve MSE close to that of MixRoute at $F = 24$ and lower MSE at $F = 360$. However, increasing the parameter count does not consistently reduce MSE. Several configurations with more parameters than MixRoute perform worse than smaller configurations. Together with the ablation results in Section 3.2.1, these findings suggest that increasing the parameter count alone is not sufficient for the Fully Shared variant to consistently match MixRoute's accuracy in the tested settings.

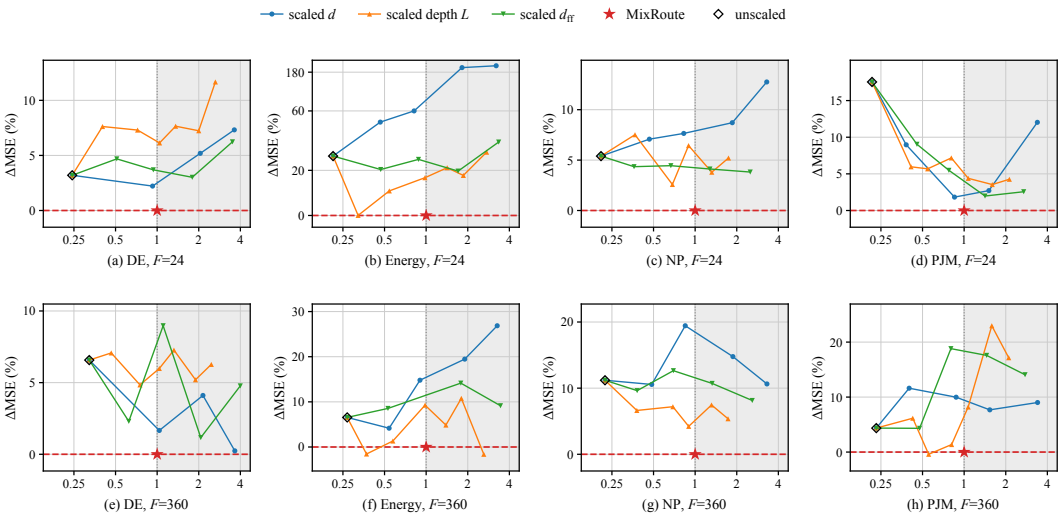


Figure 9: Scaling the Fully Shared variant of Table 3 (open diamond) along d, depth L and d_{ff}, up to $4\times$ MixRoute's parameter count. The x-axis is the parameter count relative to MixRoute on a logarithmic scale; the y-axis is the MSE relative to MixRoute (red star and dashed line; positive is worse; linear within 40%, logarithmic beyond). The shaded region covers configurations with at least as many parameters as MixRoute.

C.7 Effects of Parameter Sharing across Variables and Streams

The Fully Shared ablation shares parameters across all token types, so it does not isolate the benefit of private parameters for each exog-token. The w/o AIR ablation removes both streams, so it does not isolate the benefit of separate query projections. We compare three variants on NP, PJM, DE and Energy at $F = 24$ and $F = 360$ in the with-future setting (Table 9). The first variant shares the query, key and value projection matrices and one FFN across exog-tokens, without changing the parameter-sharing rules for endo-tokens and global tokens. The second variant uses the same sharing scheme but widens the shared FFN to match MixRoute's total parameter count. Both variants have higher average MSE than MixRoute across the four datasets at each horizon. However, the first variant has lower MSE and MAE than MixRoute on DE at $F = 24$ and lower MAE on NP at $F = 360$. On PJM at $F = 360$, its MSE and MAE match those of MixRoute at the reported precision. The third variant shares one query projection between the intrinsic and cross-type streams, leaving the rest of AIR and the training objective unchanged. This variant has higher MSE than MixRoute on every dataset at both horizons. Under the evaluated configurations, these results support private parameters for each exog-token and separate query projections.

Table 9: Effects of parameter sharing across variables and streams in the with-future setting at $F = 24$ and $F = 360$. The best results across variants are in bold.

Variant	F	NP		PJM		DE		Energy		Avg	
		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
All Exogenous Variables Share Parameters	24	0.175	0.215	0.061	0.149	**0.230**	**0.300**	0.074	0.210	0.135	0.218
	360	0.424	**0.398**	**0.121**	**0.222**	0.482	0.431	0.174	0.327	0.300	0.344
	Avg	0.300	**0.306**	0.091	0.185	0.356	0.365	0.124	0.268	0.218	0.281
Same Sharing, Same Parameter Count	24	0.170	0.214	0.064	0.150	0.243	0.304	0.070	0.205	0.137	0.218
	360	0.440	0.409	0.124	0.223	0.475	0.431	0.170	0.324	0.302	0.347
	Avg	0.305	0.311	0.094	0.186	0.359	0.368	0.120	0.265	0.220	0.282
Both Streams Share One Query Projection	24	0.172	**0.213**	0.062	0.148	0.238	0.303	0.065	0.199	0.134	0.216
	360	0.431	0.408	0.124	0.226	0.476	0.426	0.162	0.316	0.298	0.344
	Avg	0.301	0.311	0.093	0.187	0.357	0.365	0.114	0.258	0.216	0.280
MixRoute	24	**0.166**	0.215	**0.058**	**0.146**	0.233	0.302	**0.064**	**0.197**	**0.130**	**0.215**
	360	**0.412**	0.400	**0.121**	**0.222**	**0.461**	**0.421**	**0.155**	**0.308**	**0.287**	**0.338**
	Avg	**0.289**	0.308	**0.090**	**0.184**	**0.347**	**0.362**	**0.109**	**0.252**	**0.209**	**0.276**

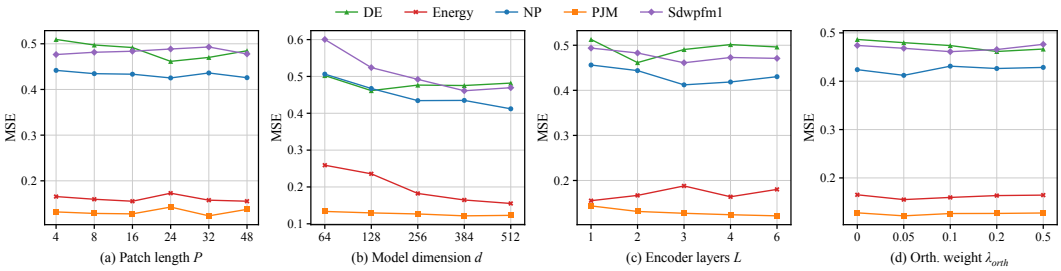


Figure 10: Parameter sensitivity of MixRoute: one axis perturbed at a time, with the remaining hyperparameters held fixed. Along the d axis, the feed-forward width scales with d. Abscissae are ordinal: grid values are equally spaced.

C.8 Parameter Sensitivity

We study the parameter sensitivity of MixRoute on five datasets. For all five datasets, we use a look-back window of 720 steps and a forecasting horizon of $F = 360$. 1) Figure 10(a) shows that the patch length with the lowest MSE varies across datasets. This variation may reflect differences in the dominant period of the endogenous series. Patch lengths between 8 and 32 perform well on four of the five datasets. Very short patches increase the token count and the cost of attention, whereas very long patches may make it harder to capture local temporal structure. 2) Figure 10(b) shows that increasing the model dimension does not consistently improve forecasting accuracy. Among the tested values, the lowest MSE is achieved with d between 256 and 512 on four of the five datasets. Excessively large dimensions may introduce redundancy in the private per-variable projections, whose parameter count grows with d^2. Smaller dimensions reduce the computational cost of these projections. 3) Figure 10(c) shows that encoders with one to three layers achieve the lowest MSE on four of the five datasets. These results suggest that deeper encoders are not always needed for accurate forecasting in the tested settings. 4) Figure 10(d) shows relatively small changes in MSE as the orthogonality weight λ_{orth} varies. Among the tested values, the best-performing weights fall between 0.05 and 0.2 on all five datasets. The relationship is not monotonic: a stronger penalty can yield higher MSE than a weaker one.

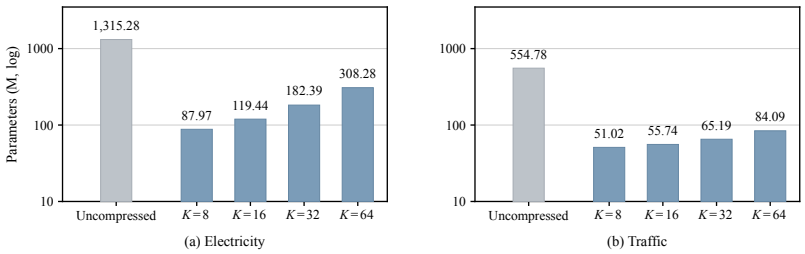


Figure 11: Trainable parameters of uncompressed and token-compressed MixRoute on Electricity (D=320) and Traffic (D=861), with future inputs and F=96. Bars use a logarithmic scale; $K \in \{8, 16, 32, 64\}$, with capacity fixed across K within each dataset.

C.9 Scaling to High-Dimensional Exogenous Variables

MixRoute's private per-variable parameters increase model size as the number of exogenous variables grows. We therefore examine token compression as a way to reduce encoder parameter costs and evaluate the resulting variant on high-dimensional auxiliary inputs. After variable-specific embedding, $\mathbf{E}_{\text{exog}} \in \mathbb{R}^{D\times d}$ is compressed into $K < D$ tokens as $\widetilde{\mathbf{E}}_{\text{exog}} = \mathbf{W}_{\text{comp}}^{\top}\mathbf{E}_{\text{exog}}$, with a learnable projection $\mathbf{W}_{\text{comp}} \in \mathbb{R}^{D\times K}$. The standard MixRoute blocks assign private attention and feed-forward parameters to each compressed token, reducing the number of exogenous parameter sets per block from D to K while retaining the D embedding projections and adding DK compression parameters.

For supplementary high-dimensional evaluation, we adapt Electricity and Traffic, two multivariate forecasting benchmarks, following DAG (Qiu et al., 2026). The last channel serves as the prediction target and the remaining channels as auxiliary variables, giving $D = 320$ and $D = 861$, respectively. Table 10 reports results for $K \in \{8, 16, 32, 64\}$ and forecasting horizons $F \in \{96, 192, 336, 720\}$, with and without future exogenous inputs. The with-future setting assumes access to future exogenous values that are not known in advance in the original forecasting task.

At $K = 32$, MixRoute has lower average MSE than both baselines on Electricity without future inputs and on Traffic in both settings. On Electricity with future inputs, however, its average MSE is 0.209, versus 0.208 for TimeXer and 0.194 for DAG. Larger K does not consistently improve accuracy. These results support using a limited number of compressed tokens in the tested settings, although the benefit varies by dataset and future-input availability. Figure 11 reports trainable parameters in the with-future setting at F=96. Capacity is fixed across K within each dataset: Electricity uses d_{model}=512, 8 heads, patch length 24 and one block; Traffic uses d_{model}=256, 4 heads and patch length 48. Across the tested K values, token compression reduces the trainable parameter count relative to uncompressed MixRoute on both datasets (Figure 11).

C.10 Varying Look-back Window

Figure 12 compares the MSE of MixRoute, DAG, TimeXer, DUET and PatchTST on DE, Energy, NP and PJM. We evaluate look-back windows $T \in \{96, 168, 336, 512, 720\}$ with the forecasting horizon fixed at $F = 360$. GCGNet is excluded because the implementation used here cannot process windows with $T < F$ when its patch length is fixed at F. MixRoute has lower MSE at $T = 720$ than at $T = 96$ on all four datasets. At $T = 720$, MixRoute achieves the lowest MSE among the five models on all four datasets, but does not always achieve the lowest MSE at shorter windows. These results suggest that MixRoute can benefit from longer input histories within the tested range.

Table 10: Forecasting results with K-token exogenous compression on Electricity (D=320) and Traffic (D=861), with and without future inputs. Avg: mean over horizons. **Red bold**: best; blue underlined: second best. Ranks use the displayed three-decimal values; ties share a rank, and second best denotes the next distinct value. The 1st Count includes all tied best results across the 20 rows, including the Avg rows.

Dataset	Setting	F	DAG		TimeXer		MixRoute							
							$K=8$		$K=16$		$K=32$		$K=64$	
			MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
Electricity	Without future	96	0.318	0.417	0.305	0.396	0.255	**0.360**	0.255	**0.360**	0.267	0.367	**0.253**	0.361
		192	0.299	0.395	0.333	0.412	0.307	0.388	0.292	**0.381**	0.292	0.383	**0.290**	0.382
		336	**0.336**	0.426	0.388	0.446	0.352	0.418	0.352	**0.415**	0.341	0.416	0.358	0.420
		720	0.396	0.469	0.456	0.493	0.382	0.459	0.391	0.464	**0.375**	**0.452**	0.444	0.487
		Avg	0.337	0.427	0.371	0.437	0.324	0.406	0.322	**0.405**	**0.319**	**0.405**	0.337	0.412
	With future	96	0.158	0.298	0.160	**0.296**	0.160	0.300	0.160	0.302	**0.156**	0.298	0.166	0.304
		192	**0.156**	**0.295**	0.186	0.320	0.199	0.334	0.185	0.324	0.180	0.320	0.188	0.330
		336	0.217	0.353	0.218	**0.346**	0.219	0.348	**0.213**	0.350	0.217	0.347	**0.213**	0.349
		720	**0.246**	**0.370**	0.267	0.386	0.282	0.400	0.278	0.399	0.282	0.402	0.274	0.399
		Avg	**0.194**	**0.329**	0.208	0.337	0.215	0.346	0.209	0.344	0.209	0.342	0.210	0.345
Traffic	Without future	96	0.164	0.240	0.183	0.279	0.141	0.213	0.146	0.218	0.140	0.211	**0.137**	**0.210**
		192	0.159	0.230	0.180	0.278	0.142	0.216	0.144	0.218	**0.141**	0.214	**0.141**	**0.213**
		336	0.157	0.235	0.180	0.284	0.146	0.224	0.147	0.226	**0.144**	0.224	**0.144**	**0.222**
		720	0.188	0.266	0.199	0.300	0.172	0.252	**0.164**	**0.244**	0.166	0.245	0.168	0.247
		Avg	0.167	0.243	0.186	0.285	0.150	0.226	0.150	0.227	**0.148**	**0.223**	**0.148**	**0.223**
	With future	96	0.119	0.187	0.135	0.233	0.106	0.185	0.108	0.184	**0.104**	**0.179**	**0.104**	**0.179**
		192	0.121	0.191	0.135	0.230	0.114	0.192	0.112	0.189	0.114	0.193	**0.109**	**0.186**
		336	0.125	0.200	0.143	0.237	0.123	0.203	0.121	0.201	**0.119**	**0.196**	0.123	0.201
		720	0.152	0.232	0.156	0.250	0.136	0.217	0.138	0.218	**0.135**	0.216	0.136	**0.214**
		Avg	0.129	0.203	0.142	0.238	0.120	0.199	0.120	0.198	**0.118**	0.196	**0.118**	**0.195**
	1st Count		4	3	0	2	0	1	2	5	**10**	5	**10**	**8**

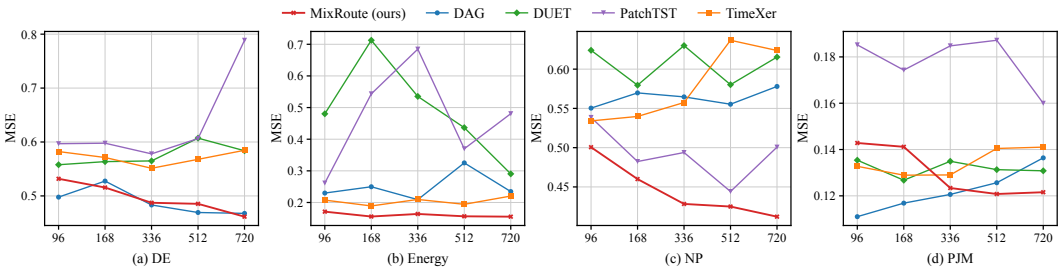
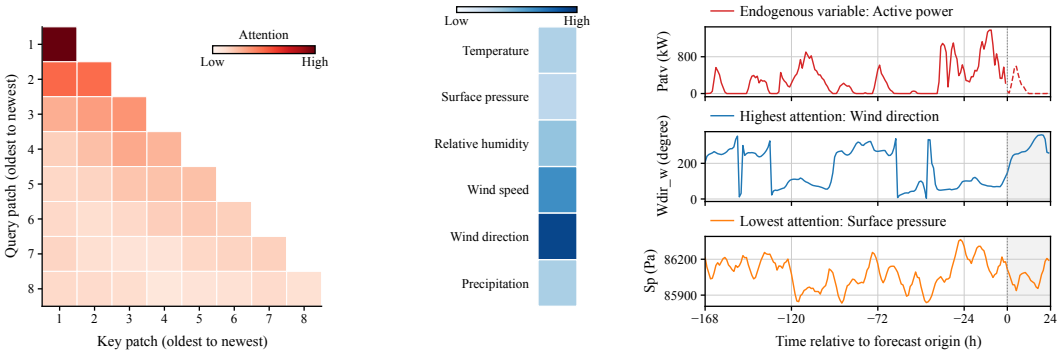


Figure 12: Forecasting performance (MSE) with varying look-back windows $T \in \{96, 168, 336, 512, 720\}$ and a fixed horizon $F = 360$.

C.11 ATTENTION VISUALIZATION



(a) Temporal attention (b) Exogenous attention (c) Endogenous and exogenous variables

Figure 13: Last-layer attention of MixRoute on Sdwpfh1 with a look-back window of 168 hours and a 24-hour forecasting horizon. (a) Attention between endogenous patches and (b) attention to exogenous variables. (c) The endogenous variable and the exogenous variables with the highest and lowest attention weights. The red dashed line shows the ground truth of the endogenous variable over the forecasting horizon.

AIR uses separate attention streams to aggregate endogenous history and read exogenous information. We examine how the intrinsic stream distributes attention across historical patches and which exogenous variables the cross-type stream emphasizes. Figure 13 shows the last-layer attention for a test window from Sdwpfh1. Panel (c) plots the endogenous series and the exogenous series with the highest and lowest attention weights to provide context for interpreting the attention maps. In panel (a), the latest endogenous query attends to both early and recent historical patches. For this query, attention is distributed across the historical window rather than concentrated on the most recent patch. In panel (b), the cross-type stream assigns the highest attention weight to wind direction, followed by wind speed, and the lowest to surface pressure.

Table 11: Results on Lisbon at $F = 10$ and $F = 30$, with and without future exogenous variables. The Avg row averages the two horizons. Red: the best, Blue: the 2nd best.

Models		**MixRoute**		DAG		GCGNet		TimeXer	
Metrics		mse	mae	mse	mae	mse	mae	mse	mae
With future	10	**0.315**	**0.409**	0.332	0.419	0.354	0.452	0.360	0.455
	30	**0.310**	**0.398**	0.327	0.417	0.363	0.458	0.351	0.454
	Avg	**0.312**	**0.404**	0.329	0.418	0.359	0.455	0.356	0.455
Without future	10	**0.313**	**0.408**	0.332	0.419	0.336	0.432	0.368	0.453
	30	**0.314**	**0.410**	0.333	0.425	0.348	0.444	0.337	0.439
	Avg	**0.314**	**0.409**	0.333	0.422	0.342	0.438	0.352	0.446

C.12 Results on a Daily Hospital Admission Task

We additionally evaluate MixRoute on the Daily Air Pollution, Meteorology, and Hospitalization Dataset for Lisbon (Relvas, 2026), hereafter Lisbon. The dataset contains 2,922 daily records from 1 January 2009 to 31 December 2016, with no missing values. Daily hospital admission counts form the endogenous series. The exogenous variables are NO2, PM10, PM2.5, minimum temperature, maximum temperature, mean temperature, relative humidity, surface pressure, wind speed, wind direction, day of week, day of year and a weekend indicator.

For this daily dataset, we follow the forecasting settings in Appendix A.3, using look-back windows of 60 and 180 days with forecasting horizons of 10 and 30 days, respectively. The split ratio is 7:1:2, as for Colbun and Rapel. Table 11 compares MixRoute, DAG, GCGNet and TimeXer at both horizons, with and without future exogenous inputs. MixRoute achieves the lowest MSE and MAE at each horizon and on average in both settings. DAG ranks second on both averaged metrics.