NMASK: A LIGHTWEIGHT NETWORK VIA MASKED ALIGNMENT FOR TIME SERIES FORECASTING WITH EXOGENOUS VARIABLES

Anonymous authors
Paper under double-blind review

ABSTRACT

Time series forecasting with exogenous variables is crucial in many real-world applications, as exogenous information often provides complementary signals beyond the target series itself. In particular, future exogenous variables, when available, offer informative priors about future dynamics and can substantially improve forecasting performance. However, existing methods often treat future exogenous variables as auxiliary features to be fused with historical observations, leading to heavy architectures and inefficient interaction modeling. In this paper, we revisit this problem from a new perspective and propose NMask, a lightweight network that explicitly leverages the inherent temporal correspondence between future exogenous variables and future prediction targets. Through a Masked Alignment mechanism, NMask associates learnable tokens with future exogenous information for future endogenous forecasting, while a Target-Conditioned Exogenous Attention module captures delayed effects of exogenous variables by modeling short-term lags and long-term temporal dependencies. Extensive experiments on multiple datasets demonstrate that NMask consistently outperforms existing forecasting methods with exogenous variables, with significantly fewer parameters and lower computational cost. Code is available at `https://anonymous.4open.science/r/NMask`.

1 INTRODUCTION

Time series forecasting is a fundamental task with widespread applications in finance (Hu et al., 2026; 2025a; Wang et al., 2024a), transportation (Wang et al., 2026c; Wu et al., 2026a; Qiu et al., 2025a; Shao et al., 2025; Xu et al., 2023), energy systems (Wang et al., 2026b; Liu et al., 2026b; Guo et al., 2015), industrial manufacturing (Cheng et al., 2026; Wang et al., 2026a; 2024b), and health (Wang et al., 2026c; Cheng et al., 2026; Wu et al., 2025a; Wei et al., 2022). Accurate forecasts enable proactive decision-making and resource allocation by estimating future values from historical observations. Recent advances (Liu et al., 2026a; Ma et al., 2026a; Ding et al., 2026; Liu et al., 2025a; Stitsyuk & Choi, 2025) in deep learning have significantly improved forecasting performance by modeling complex temporal dependencies within multivariate time series.

Beyond the endogenous (i.e., target) series itself, many real-world forecasting scenarios involve exogenous variables (i.e., covariates), which provide complementary information beyond historical endogenous observations. Based on their availability, exogenous variables can be categorized into historical and future variables. In some practical scenarios, future covariates can be obtained in advance through mature techniques. Since they directly describe external conditions during the forecasting horizon, they are particularly valuable for forecasting. For example, future weather forecasts can improve traffic prediction, while holiday schedules and promotional campaigns can enhance retail demand forecasting. Therefore, how to effectively leverage known future covariates has become an important research problem in time series forecasting (Chi et al., 2025).

Existing methods (Tu et al., 2025; Olivares et al., 2023) typically treat future covariates as additional features and incorporate them into forecasting models through feature fusion or variable interaction mechanisms. For example, TimeXer (Wang et al., 2024d) and TiDE (Das et al., 2023) use neural networks to model the dependencies between endogenous and exogenous variables, and further

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.

enhance information interaction among different variables (Hu et al., 2025b; Yu et al., 2025a; Chen et al., 2023). Although these methods achieve promising forecasting performance, their core focus remains on how to better integrate exogenous information.

However, we argue that this perspective overlooks an important property of future exogenous variables: they naturally have clear temporal correspondence with future prediction targets. As illustrated in Figure 1 (a), the future exogenous variables at each time step describe the external conditions associated with the target value at that time. Preserving this correspondence is important because different forecasting positions correspond to different external conditions, which can directly affect the variation of target values. When future covariates are mixed together during interaction, the model needs to relearn the temporal correspondence between exogenous information and prediction positions, even though this correspondence is naturally determined by the future temporal structure. Such mixing may also introduce interference among different temporal positions. Therefore, existing methods mainly focus on modeling endogenous–exogenous interactions, while rarely exploiting this temporal structure explicitly.

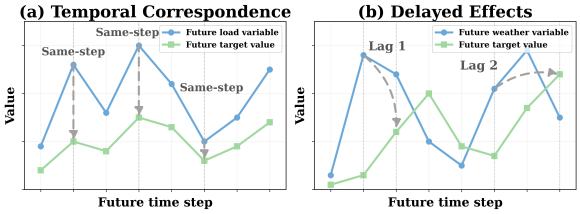


Figure 1: Illustration of temporal correspondence and delayed effects of future exogenous variables.

Moreover, the effects of future covariates are not always limited to the same time step. Temporal delays widely exist in real-world systems, where the target value at a given time is often affected by both current and previous changes in exogenous variables. As shown in Figure 1 (b), in electricity price forecasting, the price at a future time step can be influenced by weather conditions in previous hours. Therefore, effectively leveraging future covariates requires considering both their temporal correspondence with prediction positions and their delayed dependencies across time.

Based on the above observations, we propose **NMask**, a lightweight framework for forecasting with future covariates. To explicitly exploit this temporal correspondence, we introduce learnable tokens at future prediction positions to represent the unknown future targets. As illustrated in Figure 2, NMask adopts a **Masked Alignment** mechanism, which aligns these learnable tokens with the corresponding future covariates. In this way, NMask directly exploits the temporal correspondence between future covariates and prediction positions, rather than relying on complex feature interactions.

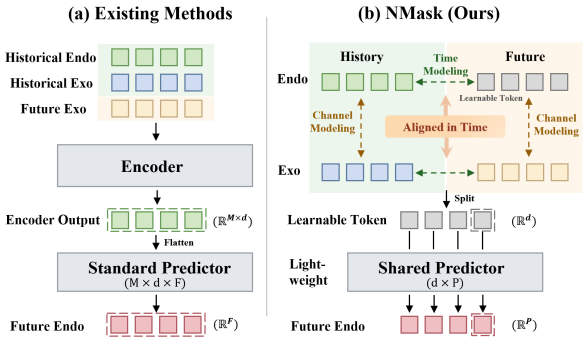


Figure 2: Comparison of existing time series forecasting methods with exogenous variables.

To further capture the delayed effects of different exogenous variables, NMask introduces a Target-Conditioned Exogenous Attention (TCEA) module. For each learnable token, TCEA independently extracts information from each exogenous variable by interacting with its corresponding neighboring historical exogenous patches and globally summarized exogenous segments. This design captures both short-term delayed responses and long-term temporal patterns from exogenous variables, while preserving the temporal alignment provided by future covariates.

Furthermore, since each learnable token is aligned with a specific future position, all tokens can share a lightweight prediction head for target patch prediction. Unlike conventional methods (Huang et al., 2026; Yue et al., 2025; Ma et al., 2025a) that map historical representations to the entire horizon, NMask reduces predictor complexity from $O(MdF)$ to $O(dP)$, where M, F, d, and P denote the number of historical patches, forecasting length, hidden dimension, and patch length, respectively.

In summary, our contributions are as follows:

- We revisit known future covariates and identify their inherent temporal correspondence with future prediction positions. NMask exploits this structure by aligning learnable future target tokens with future exogenous patches through Masked Alignment.
- We introduce a TCEA module to capture delayed effects of exogenous variables by modeling short-term lagged dependencies through interactions with individual exogenous representations.
- We design a shared patch-wise predictor that reduces predictor complexity from $O(MdF)$ to $O(dP)$, significantly reducing model complexity while maintaining forecasting performance.
- We evaluate NMask on 12 real-world datasets and demonstrate its superior forecasting performance with lower complexity and computational cost.

2 Related Work

2.1 Time Series Forecasting with Patch

Recent advances (Liu et al., 2026c; Nematirad et al., 2025; Wang et al., 2025b; Luo & Wang, 2023) in time series forecasting have been significantly influenced by patch-based representation learning, which draws inspiration from vision transformers. Instead of modeling time series at the individual timestamp level, patch-based methods (Liu et al., 2025b; Zhou et al., 2023) divide sequences into contiguous segments (patches), thereby reducing sequence length and enhancing local information.

PatchTST (Nie et al., 2023) is a representative work in this direction, showing that patching improves long-term forecasting by enabling more stable temporal representations and reducing token-level redundancy. Operating on patch-level tokens simplifies optimization and strengthens the ability of Transformers to capture long-term dependencies. Building on this idea, several works (Huang et al., 2024) explore further refinements of patch-based modeling. iTransformer (Liu et al., 2024) inverts the attention mechanism to model dependencies across variables rather than time steps, offering an alternative perspective on tokenization for multivariate forecasting. Recent studies such as PatchMLP (Tang & Zhang, 2025) and AMD (Hu et al., 2025a) have also adopted patch-based strategies to enrich sequence representations and further improve downstream task performance.

Despite their success, most patch-based methods primarily focus on improving the encoding of historical sequences, while the decoding stage remains relatively simple, typically relying on direct projection or horizon-wise forecasting heads. This design overlooks the opportunity to explicitly structure the prediction process according to the temporal semantics of the forecast horizon. In contrast, NMask explicitly models this temporal prior and introduces a shared patch-wise predictor, significantly reducing model complexity while maintaining forecasting performance.

2.2 Time Series Forecasting with Exogenous Variables

Time series forecasting with exogenous variables has been widely studied in both statistical and deep learning fields. Early models such as ARIMAX (Williams, 2001) and SARIMAX (Vagropoulos et al., 2016) incorporate external covariates into autoregressive frameworks, demonstrating the benefits of exogenous information when properly integrated with target series.

With the rise of deep learning, various architectures have been proposed to better exploit exogenous signals. Temporal Fusion Transformer (TFT) (Lim et al., 2021) integrates historical and future covariates through attention and gating mechanisms. TiDE (Das et al., 2023) adopts an MLP-based design to combine covariates with endogenous inputs. TimeXer (Wang et al., 2024d) introduces patch-wise embeddings for flexible temporal alignment, while CrossLinear (Zhou et al., 2025) models endogenous–exogenous interactions through cross-correlation embeddings. ExoTST (Tayal et al., 2024) uses cross-attention to encode exogenous variables. Recent methods, such as DAG (Qiu et al., 2026) enhances forecasting by learning and injecting temporal–channel correlation structures to better exploit exogenous variables. GCGNet (Li et al., 2026) further models variable dependencies using graph structures, but does not explicitly distinguish past and future covariates.

Although these methods differ in architecture, they mainly exploit future exogenous variables through feature interaction or aggregation, overlooking their inherent temporal prior. This may lead to unnecessary complexity while underutilizing the alignment between future covariates and targets.

In contrast, NMask explicitly align future endogenous tokens with future exogenous patches to guide the prediction of unknown future endogenous variables.

3 PRELIMINARIES

3.1 DEFINITIONS

Definition 3.1 (Time Series). Let $\mathbf{Z} \in \mathbb{R}^{N \times T}$ denotes a multivariate time series, where T represents the number of temporal observations and N denotes the number of observed variables. Each entry $Z_{i,t}$ corresponds to the value of the i-th variable at time step t. When $N = 1$, the series reduces to a univariate case; otherwise, it is multivariate.

Definition 3.2 (Endogenous Series). We define the endogenous time series $\mathbf{X}^{\text{endo}} \in \mathbb{R}^{N \times T}$ as the set of target variables whose future values need to be predicted. These variables evolve according to latent system dynamics and may be influenced by both their historical states and external signals.

Definition 3.3 (Exogenous Series). Exogenous variables refer to auxiliary covariates that are not prediction targets but provide complementary information for forecasting. We assume they are often known or can be estimated in advance. Formally, we split them into:

$$\mathbf{X}^{\text{exo}} \in \mathbb{R}^{D \times T}, \quad \mathbf{Y}^{\text{exo}} \in \mathbb{R}^{D \times F},$$

where D is the number of exogenous variates, T is the historical length, and F is the forecasting horizon. The future exogenous sequence $\mathbf{Y}^{\text{exo}}$ is assumed to be available.

3.2 PROBLEM STATEMENT

Exogenous-Aware Forecasting. Given historical endogenous observations $\mathbf{X}^{\text{endo}} \in \mathbb{R}^{N \times T}$, historical exogenous variables $\mathbf{X}^{\text{exo}} \in \mathbb{R}^{D \times T}$, and future exogenous variables $\mathbf{Y}^{\text{exo}} \in \mathbb{R}^{D \times F}$, the objective is to estimate future endogenous values over a horizon of length F. We aim to learn a parameterized mapping F_θ such that:

$$\hat{\mathbf{Y}}^{\text{endo}} = F_\theta(\mathbf{X}^{\text{endo}}, \mathbf{X}^{\text{exo}}, \mathbf{Y}^{\text{exo}}),$$

where $\hat{\mathbf{Y}}^{\text{endo}} \in \mathbb{R}^{N \times F}$ denotes the predicted future endogenous series.

4 METHODOLOGY

4.1 STRUCTURE OVERVIEW

We propose **NMask**, a lightweight forecasting framework for time series with known future exogenous variables. As illustrated in Figure 3, NMask adopts a masked alignment strategy for time series forecasting, explicitly leveraging the temporal prior between future exogenous variables and future endogenous variables by introducing learnable tokens to parameterize future endogenous segments.

NMask consists of four key components: (i) patching-based input representation, (ii) masked alignment between future exogenous variables and learnable tokens, (iii) a unified temporal-channel encoder, and (iv) a shared prediction head for efficient decoding.

4.2 PATCHING

To efficiently model long-term temporal dependencies, we first divide all inputs into non-overlapping patches with patch length P, and embed each patch into a d-dimensional representation via a linear projection. For historical endogenous variables, we obtain:

$$\mathbf{P}^{\text{endo}} = \text{PatchEmbed}(\text{Patchify}(\mathbf{X}^{\text{endo}}), \tag{1}$$

where $M = T/P$ denotes the number of historical patches and $\mathbf{P}^{\text{endo}} \in \mathbb{R}^{M \times N \times d}$.

Similarly, historical exogenous variables are transformed into:

$$\mathbf{P}^{\text{exo}} = \text{PatchEmbed}(\text{Patchify}(\mathbf{X}^{\text{exo}}), \tag{2}$$

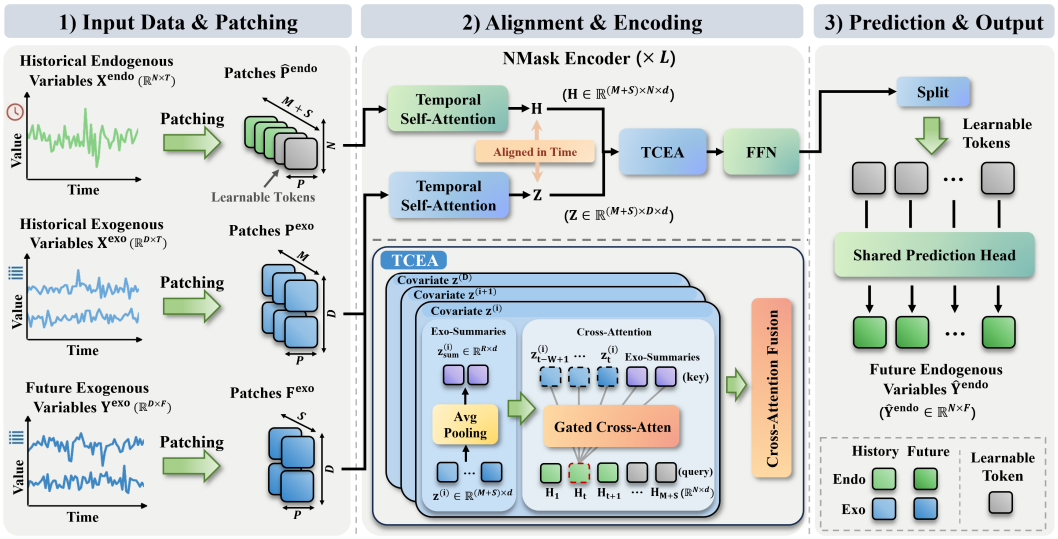


Figure 3: The architecture of NMask. The framework consists of three stages: (1) patching, (2) masked alignment with TCEA-based encoding, and (3) shared predictor for future prediction.

where $\mathbf{P}^{\text{exo}} \in \mathbb{R}^{M \times D \times d}$.

For future exogenous variables, we obtain:

$$\mathbf{F}^{\text{exo}} = \text{PatchEmbed}(\text{Patchify}(\mathbf{Y}^{\text{exo}}), \tag{3}$$

where $S = F/P$ denotes the number of future patches and $\mathbf{F}^{\text{exo}} \in \mathbb{R}^{S \times D \times d}$.

4.3 Masked Alignment

NMask reformulates forecasting with future exogenous variables as a masked prediction problem with explicit temporal alignment. Specifically, we introduce a set of learnable tokens:

$$\mathbf{E} = \{e_1, e_2, \ldots, e_S\}, \quad e_i \in \mathbb{R}^{N \times d}, \tag{4}$$

where each token represents an unknown future endogenous patch and is aligned with its corresponding future exogenous patch through a one-to-one temporal correspondence.

Unlike conventional predictors that directly generate future values from historical representations, NMask treats these tokens as latent future representations and refines them under the guidance of future exogenous variables. Specifically, the endogenous and exogenous inputs are constructed as:

$$\hat{\mathbf{P}}^{\text{endo}} = [\mathbf{P}^{\text{endo}}, \mathbf{E}], \tag{5}$$

$$\hat{\mathbf{P}}^{\text{exo}} = [\mathbf{P}^{\text{exo}}, \mathbf{F}^{\text{exo}}]. \tag{6}$$

Such alignment allows NMask to leverage the temporal prior between future exogenous variables and future endogenous segments, providing positional guidance for refining future predictions.

4.4 Encoder

After masked alignment, NMask obtains the aligned endogenous representation $\hat{\mathbf{P}}^{\text{endo}}$ and exogenous representation $\hat{\mathbf{P}}^{\text{exo}}$. The encoder employs a stack of L blocks to jointly model temporal dependencies and exogenous conditions. Each block consists of temporal self-attention, Target-Conditioned Exogenous Attention (TCEA), and a feed-forward network (FFN). For the ℓ-th layer, we first apply temporal self-attention to independently capture temporal dependencies within endogenous and exogenous variables:

$$\mathbf{H}^{(\ell)} = \text{LayerNorm}\left(\hat{\mathbf{P}}^{\text{endo}} + \text{TAttn}(\hat{\mathbf{P}}^{\text{endo}})\right), \tag{7}$$

$$\mathbf{Z}^{(\ell)} = \text{LayerNorm}\left(\hat{\mathbf{P}}^{\text{exo}} + \text{TAttn}(\hat{\mathbf{P}}^{\text{exo}})\right), \tag{8}$$

where $\mathbf{H}^{(\ell)}$ and $\mathbf{Z}^{(\ell)}$ denote the temporally enhanced representations of endogenous and exogenous variables, respectively. Since the endogenous and exogenous representations are temporally aligned, TCEA further exploits this temporal correspondence to perform exogenous-aware interactions:

$$\hat{\mathbf{H}}^{(\ell)} = \text{LayerNorm}\left(\mathbf{H}^{(\ell)} + \text{TCEA}(\mathbf{H}^{(\ell)}, \mathbf{Z}^{(\ell)})\right), \tag{9}$$

$$\mathbf{H}^{(\ell+1)} = \text{LayerNorm}\left(\hat{\mathbf{H}}^{(\ell)} + \text{FFN}(\hat{\mathbf{H}}^{(\ell)})\right). \tag{10}$$

Omitting the layer index ℓ for clarity, let $\mathbf{H}_t$ denote an endogenous patch representation and $\mathbf{z}_t^{(i)}$ denote the representation of exogenous variable i at position t. For each exogenous variable, TCEA constructs a context from local historical patches and global summaries to capture delayed effects:

$$\mathbf{z}_{\text{ctx},t}^{(i)} = [\mathbf{z}_{t-W+1}^{(i)}, \ldots, \mathbf{z}_t^{(i)}, \mathbf{z}_{\text{sum}}^{(i)}], \tag{11}$$

where W denotes the size of the historical patch window, and $\mathbf{z}_{\text{sum}}^{(i)}$ denotes R global exogenous summaries (exo-summaries) obtained by adaptive average pooling over exogenous patches.

TCEA then applies Gated Cross-Attention, where $\mathbf{H}_t$ is used as the query to perform attention over historical patches and global summaries of each exogenous variable, producing $\mathbf{c}_{\text{hist},t}^{(i)}$ and $\mathbf{c}_{\text{sum},t}^{(i)}$, respectively. The resulting contexts are combined together:

$$\mathbf{c}_t^{(i)} = \mathbf{v}_t^{(i)} + g_{\text{hist},t}^{(i)} \mathbf{c}_{\text{hist},t}^{(i)} + g_{\text{sum},t}^{(i)} \mathbf{c}_{\text{sum},t}^{(i)}, \tag{12}$$

where $\mathbf{v}_t^{(i)}$ denotes the value projection of $\mathbf{z}_t^{(i)}$, and the sigmoid gates control the contributions of historical and global contexts. The extracted information from all exogenous variables is fused as:

$$\text{TCEA}(\mathbf{H}_t, \mathbf{Z}_t) = \text{Cross-Attention Fusion}(\mathbf{c}_t^{(1)}, \ldots, \mathbf{c}_t^{(D)}). \tag{13}$$

By interacting with individual exogenous variables, TCEA captures both short-term lagged dependencies and long-term temporal patterns while maintaining the temporal alignment introduced by future covariates. The final encoder output is:

$$\mathbf{H}^{(L+1)} = \text{Encoder}(\hat{\mathbf{P}}^{\text{endo}}, \hat{\mathbf{P}}^{\text{exo}}). \tag{14}$$

4.5 Shared Prediction Head

After encoding, NMask retains only the encoded learnable tokens for forecasting. We denote these future-oriented token representations as:

$$\widetilde{\mathbf{E}} = \text{Split}\left(\mathbf{H}^{(L+1)}\right), \tag{15}$$

where $\text{Split}(\cdot)$ extracts the positions corresponding to the learnable tokens.

Each encoded token $\widetilde{\mathbf{E}}_i$ corresponds to one future endogenous patch and aggregates information from historical observations and its aligned future exogenous context. A shared prediction head $g_\phi(\cdot)$ is then applied to each token:

$$\hat{\mathbf{Y}}_i^{\text{endo}} = g_\phi\left(\widetilde{\mathbf{E}}_i\right), \quad i = 1, \ldots, S, \tag{16}$$

where $\hat{\mathbf{Y}}_i^{\text{endo}} \in \mathbb{R}^{N \times P}$ denotes the predicted endogenous patch for the i-th future segment.

The final prediction is obtained by concatenating all predicted patches along the temporal dimension:

$$\hat{\mathbf{Y}}^{\text{endo}} = \text{Concat}\left(\hat{\mathbf{Y}}_1^{\text{endo}}, \hat{\mathbf{Y}}_2^{\text{endo}}, \ldots, \hat{\mathbf{Y}}_S^{\text{endo}}\right) \in \mathbb{R}^{N \times F}. \tag{17}$$

4.6 Loss Function

NMask is optimized by minimizing the discrepancy between the predicted future endogenous series and the ground-truth future endogenous series. We adopt the mean absolute error (MAE) loss, defined as:

$$\mathcal{L} = \frac{1}{NF} \sum_{i=1}^{N} \sum_{t=1}^{F} \left| \mathbf{Y}_{i,t}^{\text{endo}} - \hat{\mathbf{Y}}_{i,t}^{\text{endo}} \right|. \tag{18}$$

The MAE loss is used due to its robustness to outliers and its effectiveness in long-horizon forecasting.

Table 1: Average results on 12 real-world datasets where future exogenous variables are available. The best performance is highlighted in **red**, and the second-best is underlined. Full results are provided in Table 5 of Appendix B.3.

Datasets	**NMask**		DAG		GCGNet		TimeXer		TFT		TiDE		DUET		CrossLinear		Amplifier		TimeKAN	
Metrics	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
NP	**0.308**	**0.320**	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
PJM	**0.085**	**0.178**	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
BE	**0.410**	**0.278**	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
FR	**0.380**	**0.216**	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
DE	**0.325**	**0.354**	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
Energy	**0.086**	**0.228**	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
Sdwpfm1	**0.375**	**0.399**	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
Sdwpfm2	**0.424**	**0.432**	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
Sdwpfh1	**0.365**	**0.411**	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
Sdwpfh2	**0.413**	**0.456**	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
Colbun	**0.097**	**0.147**	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
Rapel	**0.223**	**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
1st Count	**12**	**12**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

5 EXPERIMENT

Datasets. To ensure comprehensive and fair comparisons across different models, we conduct experiments on 12 real-world datasets with exogenous variables. In these datasets, future exogenous variables are available. We also consider settings where future exogenous variables are unavailable (see Section 5.4). Further details of the datasets are provided in Appendix A.1.

Baselines. We compare our model against 10 baselines, including 1) methods that inherently support future exogenous variables, such as DAG (Qiu et al., 2026), GCGNet (Li et al., 2026), TimeXer (Wang et al., 2024d), TiDE (Das et al., 2023), and TFT (Lim et al., 2021), and 2) methods that do not support future exogenous variables, such as DUET (Qiu et al., 2025b), CrossLinear (Zhou et al., 2025), Amplifier (Fei et al., 2025), TimeKAN (Huang et al., 2025b). For methods that do not support future covariates, we incorporate future exogenous variables via an MLP fusion approach, as detailed in Appendix A.3. For fairness, all methods use historical endogenous variables, historical exogenous variables, and future exogenous variables as inputs.

Implementation Details. To be consistent with previous works, we adopt Mean Squared Error (MSE) and Mean Absolute Error (MAE) as evaluation metrics. We conduct both long-term and short-term forecasting experiments. For the Colbún and Rapel datasets, we set the lookback window to 60 with a prediction horizon of 10 for short-term forecasting, and a lookback window of 180 with a prediction horizon of 30 for long-term forecasting. For the other datasets, we use a lookback window of 168 with a prediction horizon of 24 for short-term forecasting, and a lookback window of 720 with a prediction horizon of 360 for long-term forecasting.

We utilize the TFB (Qiu et al., 2024) code repository for standardized evaluation, and all baseline results are obtained using TFB. Following TFB (Qiu et al., 2024) and TSFM-Bench (Li et al., 2025), we do not apply the "Drop Last" trick to ensure fair comparisons. Further implementation details are provided in Appendix A.2.

5.1 MAIN RESULTS

Comprehensive forecasting results are presented in Table 1 to demonstrate the performance of NMask on the datasets with exogenous variables. Across both long-term and short-term forecasting settings on 12 real-world datasets, NMask consistently outperforms a wide range of forecasters with diverse architectures. Specifically, as shown in Table 1, NMask achieves the best performance on all datasets. The results highlight NMask's clear advantage over models specifically designed for time series forecasting with exogenous variables that adopt a two-step modeling strategy, such as TimeXer and TFT, as well as recent state-of-the-art forecasters adapted to this setting, including DUET and Amplifier. By explicitly leveraging the temporal structure between future exogenous

Table 2: Ablation studies for NMask. The best results are highlighted in **bold**.

Datasets	NP		PJM		DE		Energy		Average	
Metrics	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
(a) Replace Masked Alignment	0.495	0.461	0.130	0.227	0.446	0.419	0.164	0.319	0.309	0.357
(b) Remove TCEA	0.456	0.428	0.115	0.214	0.416	0.404	0.126	0.281	0.278	0.332
(c) Use All History	0.567	0.481	0.128	0.231	0.470	0.430	0.187	0.340	0.338	0.371
(d) Use All Future	0.482	0.461	0.112	0.211	0.440	0.417	0.142	0.294	0.294	0.346
(e) Use All Sequence	0.456	0.427	0.118	0.225	0.451	0.418	0.163	0.317	0.297	0.347
NMask (ours)	**0.437**	**0.419**	**0.110**	**0.209**	**0.400**	**0.395**	**0.114**	**0.268**	**0.265**	**0.323**

variables and future endogenous segments rather than treating them as auxiliary features for fusion, NMask achieves consistently stronger performance and surpasses these competitive baselines on nearly all evaluation metrics. These results further support the effectiveness of NMask for leveraging future exogenous information and capturing delayed effects in time series forecasting.

5.2 Model Analyses

5.3 Ablation Study

We conduct ablation studies to evaluate the key designs of NMask, as shown in Table 2.

(a) Replace Masked Alignment. We replace the proposed Masked Alignment with an MLP fusion approach (detailed in Appendix A.3). Instead of explicitly aligning learnable future endogenous tokens with future exogenous variables, the MLP-based method directly fuses future exogenous variables with endogenous representations. The performance degradation demonstrates the importance of using future temporal prior for Masked Alignment.

(b) Remove TCEA. We remove the Target-Conditioned Exogenous Attention (TCEA) module and directly fuse each endogenous patch with its corresponding exogenous patch. The performance degradation demonstrates the importance of TCEA in capturing variable-specific delayed effects and long-term dependencies of exogenous variables.

(c)–(e) Prediction strategies. In NMask, the shared predictor enables each learnable future endogenous token to predict a single future endogenous patch, maintaining an explicit correspondence between the token and its target patch. We compare it with three alternatives: **(c)** using all historical endogenous hidden representations to predict the future sequence, **(d)** using all learnable future endogenous tokens jointly for prediction, and **(e)** combining historical hidden representations and future tokens to predict the entire future sequence.

As shown in Table 2, NMask consistently achieves the best performance across all datasets. These results validate that both the future temporal prior leveraged by Masked Alignment and the shared token-to-patch prediction strategy are effective.

5.4 Forecasting without Future Exogenous Variables

Considering that future exogenous variables may not always be available in real-world scenarios, we further evaluate the robustness and applicability of NMask using only historical endogenous and exogenous variables, with results reported in Table 3. When future exogenous variables $\mathbf{Y}^{\text{exo}}$ are unavailable, NMask replaces them with learnable future exogenous tokens while preserving the remaining architecture unchanged. An auxiliary loss and an exogenous channel attention module are introduced to enable these tokens to learn informative representations when future covariates are unavailable. For all baseline methods, only historical variables are provided as inputs to ensure a fair comparison. Several baselines, including DAG and GCGNet, are capable of operating without future exogenous variables. DUET models inter-variable dependencies and therefore can also naturally utilize historical endogenous and exogenous variables. The experimental results show that NMask maintains competitive performance even without future exogenous variables.

Table 3: Average results on 12 real-world datasets where future exogenous variables are not available. The best performance is highlighted in **red**, and the second-best is underlined. Full results are provided in Table 6 of Appendix B.3.

Datasets	**NMask**		DAG		GCGNet		TimeXer		TFT		TiDE		DUET		CrossLinear		Amplifier		TimeKAN	
Metrics	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
NP	0.428	**0.366**	**0.419**	0.380	0.430	0.381	0.440	0.383	0.647	0.488	0.466	0.415	0.443	0.383	0.451	0.394	0.520	0.448	0.484	0.436
PJM	**0.123**	**0.210**	0.126	0.210	0.131	0.216	0.141	0.229	0.200	0.270	0.158	0.252	0.140	0.224	0.147	0.241	0.152	0.245	0.175	0.270
BE	**0.457**	**0.290**	0.462	0.297	0.459	0.306	0.477	0.301	0.563	0.354	0.547	0.349	0.473	0.305	0.477	0.300	0.503	0.325	0.488	0.315
FR	**0.431**	**0.239**	0.435	0.249	0.435	0.246	0.454	0.247	0.535	0.288	0.493	0.289	0.467	0.262	0.476	0.257	0.494	0.286	0.491	0.277
DE	**0.593**	**0.477**	0.603	0.487	0.614	0.501	0.660	0.507	0.684	0.515	0.643	0.519	0.659	0.513	0.635	0.508	0.712	0.549	0.669	0.526
Energy	0.159	0.308	0.151	0.300	0.154	0.305	0.172	0.326	0.375	0.480	0.163	0.312	0.160	0.308	0.201	0.336	0.180	0.327	**0.148**	**0.296**
Sdwpfm1	0.705	**0.570**	**0.689**	0.604	0.711	0.596	0.702	0.609	0.984	0.728	0.712	0.633	0.724	0.585	0.810	0.694	0.843	0.686	0.726	0.672
Sdwpfm2	**0.778**	**0.615**	0.786	0.657	0.807	0.643	0.803	0.653	1.042	0.766	0.816	0.683	0.818	0.641	0.800	0.687	0.942	0.732	0.836	0.701
Sdwpfh1	0.767	**0.622**	**0.734**	0.643	0.741	0.628	0.746	0.643	0.793	0.697	0.805	0.698	0.780	0.644	0.768	0.702	1.100	0.820	0.804	0.721
Sdwpfh2	**0.855**	**0.669**	0.870	0.695	0.886	0.701	0.891	0.719	0.926	0.746	0.920	0.752	1.007	0.715	0.956	0.774	0.980	0.768	0.941	0.783
Colbun	**0.110**	**0.150**	0.117	0.155	0.119	0.165	0.132	0.218	0.556	0.387	0.187	0.242	0.146	0.197	0.129	0.195	0.152	0.211	0.201	0.253
Rapel	0.258	**0.274**	0.261	0.281	0.265	0.284	0.344	0.362	0.308	0.334	0.324	0.357	0.305	0.304	**0.241**	0.299	0.337	0.348	0.342	0.338
1st Count	**7**	**11**	3	0	0	0	0	0	0	0	0	0	0	0	1	0	0	0	1	1

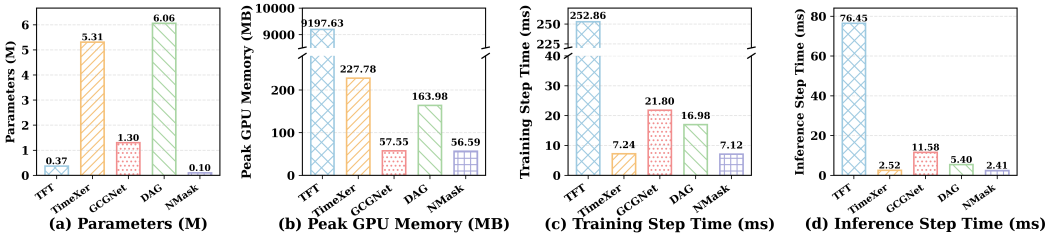


Figure 4: Efficiency comparison of different forecasting models on the NP dataset.

5.5 Efficiency Analysis

We provide an efficiency comparison on the NP dataset with an input length of 360, as reported in Figure 4. The comparison considers four aspects, including the number of parameters, GPU memory consumption, training step time, and inference step time, providing a comprehensive evaluation of computational cost and runtime efficiency. Overall, NMask achieves superior efficiency compared with all baselines. It requires fewer parameters and less computational overhead while maintaining competitive forecasting performance. This efficiency mainly benefits from explicitly leveraging the temporal correspondence between future exogenous variables and future prediction positions, which simplifies the model design, together with the lightweight shared predictor that avoids the need for large-scale forecasting heads. In contrast, other methods generally introduce more complex architectures and require higher computational costs.

6 Conclusion

In this paper, we revisit forecasting with known future covariates from a new perspective and propose **NMask**, a lightweight framework that explicitly leverages the temporal correspondence between future exogenous variables and future prediction positions through Masked Alignment. By aligning learnable future endogenous tokens with future exogenous patches, NMask effectively exploits future temporal priors while avoiding complex feature interactions. In addition, the proposed TCEA module captures delayed effects of exogenous variables, and the shared patch-wise predictor improves parameter efficiency. Extensive experiments on 12 real-world datasets demonstrate that NMask achieves competitive forecasting performance with substantially lower model complexity. In the future, we will explore more challenging scenarios with uncertain future covariates and investigate more adaptive mechanisms for modeling complex exogenous dependencies.

ETHICS STATEMENT

Our work exclusively uses publicly available benchmark datasets that contain no personally identifiable information. No human subjects are involved in this research.

AI USE STATEMENT

We used ChatGPT to assist in polishing the manuscript and improving the clarity of the writing. All AI-assisted content was carefully reviewed by the authors. All methodological designs and experimental analyses were conducted independently by the authors, and the use of AI was limited to auxiliary purposes. The authors take full responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

We promise that all experimental results can be reproduced. We have released our model code in an anonymous repository: `https://anonymous.4open.science/r/NMask`.

REFERENCES

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

Fang Cheng, Hui Liu, and Xinwei Lv. MetaGNSDformer: Meta-learning enhanced gated non-stationary informer with frequency-aware attention for point-interval remaining useful life prediction of lithium-ion batteries. *Advanced Engineering Informatics*, 69:103798, 2026.

Ce Chi, Xing Wang, Kexin Yang, Zhiyan Song, Di Jin, Lin Zhu, Chao Deng, and Junlan Feng. InjectTST: Injecting global information into independent channels for long time series forecasting. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pp. 1–5, 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.

Kuiye Ding, Fanda Fan, Chunyi Hou, Zheya Wang, Lei Wang, Zhengxin Yang, and Jianfeng Zhan. TimeMosaic: Temporal heterogeneity guided time series forecasting via adaptive granularity patch and segment-wise decoding. In *AAAI Conference on Artificial Intelligence*, volume 40, pp. 20790–20798, 2026.

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

Chenjuan Guo, Bin Yang, Ove Andersen, Christian Jensen, and Kristian Torp. EcoMark 2.0: empowering eco-routing with vehicular environmental models and actual vehicle fuel consumption data. *GeoInformatica*, 19:567–599, 2015.

Hans Hersbach, Bill Bell, Paul Berrisford, Shoji Hirahara, András Horányi, Joaquín Muñoz-Sabater, Julien Nicolas, Carole Peubey, Raluca Radu, Dinand Schepers, and et al. The era5 global reanalysis. *Quarterly Journal of the Royal Meteorological Society*, 146:1999 – 2049, 2020.

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

Yifan Hu, Guibin Zhang, Peiyuan Liu, Disen Lan, Naiqi Li, Dawei Cheng, Tao Dai, Shu-Tao Xia, and Shirui Pan. TimeFilter: Patch-specific spatial-temporal graph filtration for time series forecasting. In *International Conference on Machine Learning*, volume 267, pp. 24893–24911, 2025b.

Yifan Hu, Yuante Li, Peiyuan Liu, Yuxia Zhu, Naiqi Li, Tao Dai, Shu tao Xia, Dawei Cheng, and Changjun Jiang. Fintsb: A comprehensive and practical benchmark for financial time series forecasting. *arXiv preprint arXiv:2502.18834*, 2025a.

Yifan Hu, Jie Yang, Tian Zhou, Peiyuan Liu, Yujin Tang, Rong Jin, and Liang Sun. Bridging past and future: Distribution-aware alignment for time series forecasting. In *International Conference on Learning Representations*, pp. 126148–126172, 2026.

Qihe Huang, Lei Shen, Ruixin Zhang, Jiahuan Cheng, Shouhong Ding, Zhengyang Zhou, and Yang Wang. HDMixer: Hierarchical dependency with extendable patch for multivariate time series forecasting. In *AAAI Conference on Artificial Intelligence*, volume 38, pp. 12608–12616, 2024.

Qihe Huang, Zhengyang Zhou, Kuo Yang, Zhongchao Yi, Xu Wang, and Yang Wang. TimeBase: The power of minimalism in efficient long-term time series forecasting. In *International Conference on Machine Learning*, pp. 26227–26246, 2025a.

Qihe Huang, Zhengyang Zhou, Yangze Li, Jiaming Ma, Kuo Yang, Binwu Wang, Xu Wang, and Yang Wang. STORM: Synergistic cross-scale spatio-temporal modeling for weather forecasting. In *International Conference on Learning Representations*, 2026.

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*, 2025b.

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.

Zhe Li, Xiangfei Qiu, Peng Chen, Yihang Wang, Hanyin Cheng, Yang Shu, Jilin Hu, Chenjuan Guo, Aoying Zhou, Christian S. Jensen, and Bin Yang. TSFM-Bench: A comprehensive and unified benchmark of foundation models for time series forecasting. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, pp. 5595–5606, 2025.

Zhengyu Li, Xiangfei Qiu, Yuhan Zhu, Xingjian Wu, Jilin Hu, Guo Guo, and Bin Yang. GCGNet: Graph-consistent generative network for time series forecasting with exogenous variables. In *International Conference on Learning Representations*, 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.

Chenxi Liu, Qianxiong Xu, Hao Miao, Sun Yang, Lingzheng Zhang, Cheng Long, Ziyue Li, and Rui Zhao. TimeCMA: Towards llm-empowered multivariate time series forecasting via cross-modality alignment. In *AAAI Conference on Artificial Intelligence*, volume 39, pp. 18780–18788, 2025a.

Hao Liu, Zhang xiaoxing, Chun Yang, and Xiaobin Zhu. Semantic-enhanced time-series forecasting via large language models. In *International Conference on Learning Representations*, 2026a.

Peiyuan Liu, Hang Guo, Tao Dai, Naiqi Li, Jigang Bao, Xudong Ren, Yong Jiang, and Shu-Tao Xia. CALF: Aligning llms for time series forecasting via cross-modal fine-tuning. In *AAAI Conference on Artificial Intelligence*, volume 39, pp. 18915–18923, 2025b.

Xvyuan Liu, Xiangfei Qiu, Hanyin Cheng, Xingjian Wu, Chenjuan Guo, Bin Yang, and Jilin Hu. ASTGI: Adaptive spatio-temporal graph interactions for irregular multivariate time series forecasting. In *International Conference on Learning Representations*, 2026b.

Xvyuan Liu, Xiangfei Qiu, Xingjian Wu, Zhengyu Li, Chenjuan Guo, Jilin Hu, and Bin Yang. Rethinking irregular time series forecasting: A simple yet effective baseline. In *AAAI Conference on Artificial Intelligence*, 2026c.

Yong Liu, Tengge Hu, Haoran Zhang, Haixu Wu, Shiyu Wang, Lintao Ma, and Mingsheng Long. iTransformer: Inverted transformers are effective for time series forecasting. In *International Conference on Learning Representations*, 2024.

Donghao Luo and Xue Wang. Cross-LKTCN: Modern convolution utilizing cross-variable dependency for multivariate time series forecasting. *arXiv preprint arXiv:2306.02326*, 2023.

Jiaming Ma, Binwu Wang, Guanjun Wang, Kuo Yang, Zhengyang Zhou, Pengkun Wang, Xu Wang, and Yang Wang. Less but more: Linear adaptive graph learning empowering spatiotemporal forecasting. In *Advances in Neural Information Processing Systems*, volume 38, pp. 58382–58418, 2025a.

Jiaming Ma, Qihe Huang, Haofeng Ma, Guanjun Wang, Sheng Huang, Zhengyang Zhou, Pengkun Wang, Xu Wang, Binwu Wang, and Yang Wang. PHAT: Modeling period heterogeneity for multivariate time series forecasting. In *International Conference on Learning Representations*, 2026a.

Yue Ma, Yingqing He, Xiaodong Cun, Xintao Wang, Siran Chen, Ying Shan, Xiu Li, and Qifeng Chen. Follow your pose: Pose-guided text-to-video generation using pose-free videos. In *AAAI Conference on Artificial Intelligence*, volume 38, pp. 4117–4125, 2024a.

Yue Ma, Hongyu Liu, Hongfa Wang, Heng Pan, Yingqing He, Junkun Yuan, Ailing Zeng, Chengfei Cai, Heung-Yeung Shum, Wei Liu, and Qifeng Chen. Follow-your-emoji: Fine-controllable and expressive freestyle portrait animation. In *SIGGRAPH Asia 2024 Conference Papers*, pp. 1–12, 2024b.

Yue Ma, Yingqing He, Hongfa Wang, Andong Wang, Chenyang Qi, Chengfei Cai, Xiu Li, Zhifeng Li, Heung-Yeung Shum, Wei Liu, and Qifeng Chen. Follow-your-click: Open-domain regional image animation via short prompts. In *AAAI Conference on Artificial Intelligence*, volume 39, pp. 6018–6026, 2025b.

Yue Ma, Zexuan Yan, Hongyu Liu, Hongfa Wang, Heng Pan, Yingqing He, Junkun Yuan, Ailing Zeng, Chengfei Cai, Heung-Yeung Shum, Zhifeng Li, Wei Liu, Linfeng Zhang, and Qifeng Chen. Follow-your-emoji-faster: Towards efficient, fine-controllable, and expressive freestyle portrait animation. *arXiv preprint arXiv:2509.16630*, 2025c.

Yue Ma, Kunyu Feng, Zhongyuan Hu, Xinyu Wang, Yucheng Wang, Mingzhe Zheng, Bingyuan Wang, Qinghe Wang, Xuanhua He, Hongfa Wang, Chenyang Zhu, Hongyu Liu, Yingqing He, and et al. Controllable video generation: A survey. *arXiv preprint arXiv:2507.16869*, 2026b.

Yue Ma, Yulong Liu, Qiyuan Zhu, Ayden Yang, Kunyu Feng, Xinhua Zhang, Zexuan Yan, Zhifeng Li, Sirui Han, Chenyang Qi, and Qifeng Chen. Follow-your-motion: Video motion transfer via efficient spatial-temporal decoupled finetuning. *arXiv preprint arXiv:2506.05207*, 2026c.

Reza Nematirad, Anil Pahwa, and Balasubramaniam Natarajan. Times2D: Multi-period decomposition and derivative mapping for general time series forecasting. In *AAAI Conference on Artificial Intelligence*, volume 39, pp. 19651–19658, 2025.

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.

Kin G. Olivares, Cristian Challu, Grzegorz Marcjasz, Rafał Weron, and Artur Dubrawski. Neural basis expansion analysis with exogenous variables: Forecasting electricity prices with nbeatsx. *International Journal of Forecasting*, 39(2):884–900, 2023.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, and DeVito. Pytorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, volume 32, pp. 8024–8035, 2019.

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

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

Xiangfei Qiu, Xingjian Wu, Yan Lin, Chenjuan Guo, Jilin Hu, and Bin Yang. DUET: Dual clustering enhanced multivariate time series forecasting. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.1*, pp. 1185–1196, 2025b.

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

Zongjiang Shang, Ling Chen, Binqing Wu, and Dongliang Cui. Ada-MSHyper: Adaptive multi-scale hypergraph transformer for time series forecasting. In *Advances in Neural Information Processing Systems*, volume 37, pp. 33310–33337, 2024.

Wei Shao, Ziquan Fang, Lu Chen, and Yunjun Gao. Towards trajectory anomaly detection: a fine-grained and noise-resilient framework. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 2490–2501, 2025.

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

Peiwang Tang and Weitai Zhang. Unlocking the power of patch: Patch-based mlp for long-term time series forecasting. In *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 *IEEE International Conference on Data Mining*, pp. 857–862, 2024.

Shihao Tu, Yang Yang, Wenyue Ding, Yicheng Lu, Qingkai Ren, Yupeng Zhang, and Yin Zhang. ASTNet: Asynchronous spatio-temporal network for large-scale chemical sensor forecasting. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, pp. 4913–4924, 2025.

Stylianos I. Vagropoulos, G. I. Chouliaras, E. G. Kardakos, C. K. Simoglou, and A. G. Bakirtzis. Comparison of sarimax, sarima, modified sarima and ann-based models for short-term pv generation forecasting. In *IEEE International Energy Conference*, pp. 1–6, 2016.

Hao Wang, Licheng Pan, Zhichao Chen, Degui Yang, Sen Zhang, Yifei Yang, Xinggao Liu, Haoxuan Li, and Dacheng Tao. FreDF: Learning to forecast in the frequency domain. In *International Conference on Learning Representations*, 2024a.

Hao Wang, Zhiyu Wang, Yunlong Niu, Zhaoran Liu, Haozhe Li, Yilin Liao, Yuxin Huang, and Xinggao Liu. An accurate and interpretable framework for trustworthy process monitoring. *IEEE Transactions on Artificial Intelligence*, 5(5):2241–2252, 2024b.

Hao Wang, zhengnan li, Haoxuan Li, Xu Chen, Mingming Gong, BinChen, and Zhichao Chen. Optimal transport for time series imputation. In *International Conference on Learning Representations*, pp. 828–852, 2025a.

Hao Wang, Zhichao Chen, Licheng Pan, Xiaoyu Jiang, Yichen Song, Qunshan He, and Xinggao Liu. DeepFilter: A transformer-style framework for accurate and efficient process monitoring. *IEEE Transactions on Artificial Intelligence*, 7:3979–3989, 2026a.

Hao Wang, Licheng Pan, Zhichao Chen, Xu Chen, Qingyang Dai, Lei Wang, Haoxuan Li, and Zhouchen Lin. Time-o1: Time-series forecasting needs transformed label alignment. In *Advances in Neural Information Processing Systems*, volume 38, pp. 8660–8690, 2026b.

Hao Wang, Licheng Pan, Yuan Lu, Zi Ciu Chan, Tianqiao Liu, Shuting He, Zhixuan Chu, Qingsong Wen, Haoxuan Li, and Zhouchen Lin. Quadratic direct forecast for training multi-step time-series forecast models. In *International Conference on Learning Representations*, 2026c.

Hao Wang, Licheng Pan, Yuan Lu, Zhixuan Chu, Xiaoxi Li, Shuting He, Zhichao Chen, Haoxuan Li, Qingsong Wen, and Zhouchen Lin. DistDF: Time-series forecasting needs joint-distribution wasserstein alignment. In *International Conference on Learning Representations*, 2026c.

Shiyu Wang, Haixu Wu, Xiaoming Shi, Tengge Hu, Huakun Luo, Lintao Ma, James Y. Zhang, and Jun Zhou. TimeMixer: Decomposable multiscale mixing for time series forecasting. In *International Conference on Learning Representations*, 2024c.

Shiyu Wang, Jiawei LI, Xiaoming Shi, Zhou Ye, Baichuan Mo, Wenze Lin, Ju Shengtong, Zhixuan Chu, and Ming Jin. TimeMixer++: A general time series pattern machine for universal predictive analysis. In *International Conference on Learning Representations*, 2025b.

Yuxuan Wang, Haixu Wu, Jiaxiang Dong, Guo Qin, Haoran Zhang, Yong Liu, Yunzhong Qiu, Jianmin Wang, and Mingsheng Long. TimeXer: Empowering transformers for time series forecasting with exogenous variables. In *Advances in Neural Information Processing Systems*, volume 37, pp. 469–498, 2024d.

Kaimin Wei, Tianqi Li, Feiran Huang, Jinpeng Chen, and Zefan He. Cancer classification with data augmentation based on generative adversarial networks. *Frontiers of Computer Science*, 16:1–11, 2022.

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

Binqing Wu, Zongjiang Shang, Jianlong Huang, and Ling Chen. MillGNN: Learning multi-scale lead-lag dependencies for multi-variate time series forecasting. *ACM International Conference on Information and Knowledge Management*, 2025a.

Binqing Wu, Zongjiang Shang, Shiyu Liu, Jianlong Huang, Jiahui Xu, and Ling Chen. AirDDE: Multifactor neural delay differential equations for air quality forecasting. In *AAAI Conference on Artificial Intelligence*, volume 40, pp. 15941–15949, 2026a.

Xingjian Wu, Xiangfei Qiu, Zhengyu Li, Yihang Wang, Jilin Hu, Chenjuan Guo, Hui Xiong, and Bin Yang. Catch: Channel-aware multivariate time series anomaly detection via frequency patching. In *International Conference on Learning Representations*, pp. 17017–17045, 2025b.

Xingjian Wu, Jianxin Jin, Wanghui Qiu, Peng Chen, Yang Shu, Bin Yang, and Chenjuan Guo. Aurora: Towards universal generative multimodal time series forecasting. In *International Conference on Learning Representations*, 2026b.

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

Ronghui Xu, Meng Chen, Yongshun Gong, Yang Liu, Xiaohui Yu, and Liqiang Nie. TME: Tree-guided multi-task embedding learning towards semantic venue annotation. *ACM Transactions on Information Systems*, 41(4):1–24, 2023.

Sean Bin Yang, Jilin Hu, Chenjuan Guo, Bin Yang, and Christian S. Jensen. LightPath: Lightweight and scalable path representation learning. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 2999–3010, 2023.

Chengqing Yu, Fei Wang, Zezhi Shao, Tangwen Qian, Zhao Zhang, Wei Wei, Zhulin An, Qi Wang, and Yongjun Xu. GinAR+: A robust end-to-end framework for multivariate time series forecasting with missing values. *IEEE Transactions on Knowledge and Data Engineering*, 37(8): 4635–4648, 2025a.

Chengqing Yu, Fei Wang, Chuanguang Yang, Zezhi Shao, Tao Sun, Tangwen Qian, Wei Wei, Zhulin An, and Yongjun Xu. Merlin: Multi-view representation learning for robust multivariate time series forecasting with unfixed missing rates. In *ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, pp. 3633–3644, 2025b.

Wenzhen Yue, Yong Liu, Hao Wang, Haoxuan Li, Xianghua Ying, Ruohao Guo, Bowei Xing, and Ji Shi. OLinear: A linear model for time series forecasting in orthogonally transformed domain. In *Advances in Neural Information Processing Systems*, volume 38, pp. 60573–60632, 2025.

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 *ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2*, 2025.

Tian Zhou, Peisong Niu, xue wang, Liang Sun, and Rong Jin. One Fits All: Power general time series analysis by pretrained lm. In *Advances in Neural Information Processing Systems*, volume 36, pp. 43322–43355, 2023.

A Experimental Details

A.1 Datasets

Table 4: Statistics of datasets. Ex. and En. are abbreviations for the Exogenous variable and Endogenous variable, respectively.

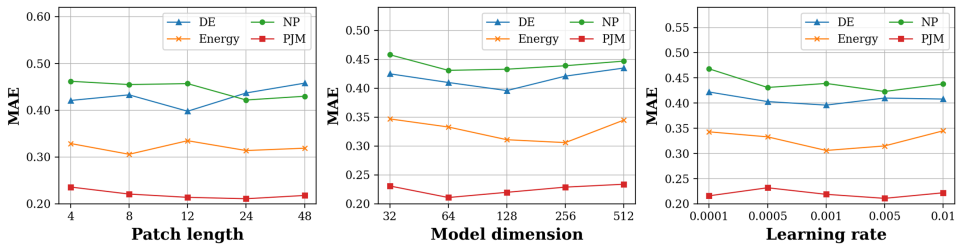
Dataset	#Num	Sampling Frequency	Ex. Descriptions	En. Descriptions	Lengths	Split
NP	2	1 Hour	Grid Load, Wind Power	Nord Pool Electricity Price	52,416	7:1:2
PJM	2	1 Hour	System Load, SyZonal COMED load	Pennsylvania-New Jersey-Maryland Electricity Price	52,416	7:1:2
BE	2	1 Hour	Generation, System Load	Belgium's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

We evaluate our method on 12 real-world datasets involving exogenous variables whose future values are either approximately available in advance or can be estimated with high accuracy. The benchmark consists of five electricity price forecasting (EPF) datasets from Wang et al. (2024c) and Lago et al. (2021), including NP, PJM, BE, FR, and DE, together with seven datasets introduced by DAG (Qiu et al., 2026), namely Energy, Colbún, Rapel, Sdwpfh1, Sdwpfh2, Sdwpfm1, and Sdwpfm2.

The five EPF datasets cover electricity markets in different European and U.S. regions. In **NP**, the prediction target is the hourly Nord Pool electricity price, while grid load and wind power forecasts are provided as exogenous information. **PJM** contains electricity prices from the COMED area of the Pennsylvania–New Jersey–Maryland Interconnection, accompanied by forecasts of system load and COMED zonal load. For **BE**, hourly Belgian electricity prices serve as the endogenous variable, with Belgian load forecasts and French generation forecasts used as exogenous inputs. **FR** considers electricity prices in the French market, together with generation and load forecasts. Similarly, **DE** focuses on hourly German electricity prices and incorporates forecasts of Amprion zonal load, wind generation, and solar generation as external variables.

The remaining datasets span power generation, hydropower, and wind power applications. **Energy** contains hourly power generation measurements from Chile, where thermoelectric generation is used as the forecasting target and battery storage, wind, hydroelectric, and solar generation provide exogenous information. **Colbún** and **Rapel** are daily hydropower datasets, in which water level is predicted using precipitation and tributary inflow as external variables. **Sdwpfh1**, **Sdwpfh2**, **Sdwpfm1**, and **Sdwpfm2** are collected from the Longyuan wind farm. Their prediction target is active power output (Patv), while meteorological variables obtained from ERA5 (Hersbach et al., 2020), including temperature, surface pressure, relative humidity, wind speed, wind direction, and total precipitation, are used as exogenous features.

Evaluating models across datasets from multiple domains has become a common practice in modern deep learning, as it provides a more reliable assessment of model generalization under heterogeneous data distributions and application scenarios (Ma et al., 2025c; 2024a;b; Wu et al., 2024; Yang et al., 2023). This protocol has been widely adopted in natural language processing and computer vision (Ma et al., 2026c;b; 2025b), as well as in various time-series tasks, including forecasting (Wang et al., 2026c; Shang et al., 2024), anomaly detection (Wu et al., 2025b), imputation (Wang et al., 2025a), irregular time-series modeling (Liu et al., 2026c; Yu et al., 2025b), long-term modeling (Liu et al., 2026c), and general-purpose time-series modeling (Wu et al., 2026b). Following this widely used evaluation protocol, our benchmark covers multiple real-world domains with distinct temporal patterns, sampling frequencies, and exogenous-variable configurations, enabling a comprehensive examination of model performance across diverse forecasting scenarios.



(a) Patch length (b) Model dimension (c) Learning rate

Figure 5: Sensitivity analysis of key hyperparameters for NMask with a prediction horizon of 360.

A.2 Implementation Details

To keep consistent with previous works, we adopt Mean Squared Error (MSE) and Mean Absolute Error (MAE) as evaluation metrics.

We conduct both long-term and short-term forecasting experiments. For the Colbún and Rapel datasets, we set the lookback window to 60 with a prediction horizon of 10 for short-term forecasting, and a lookback window of 180 with a prediction horizon of 30 for long-term forecasting. For the other datasets, we use a lookback window of 168 with a prediction horizon of 24 for short-term forecasting, and a lookback window of 720 with a prediction horizon of 360 for long-term forecasting.

We utilize the TFB (Qiu et al., 2024) code repository for unified evaluation, with all baseline results also derived from TFB. Following the settings in TFB and TSFM-Bench (Li et al., 2025), we do not apply the "Drop Last" trick to ensure a fair comparison.

All experiments of NMask are conducted using PyTorch (Paszke et al., 2019) in Python 3.11 and executed on an NVIDIA Tesla-A800 GPU. The training process is guided by the L1 loss function and employs the ADAM optimizer. The initial batch size is set to 64, with the flexibility to halve it (down to a minimum of 8) in case of an Out-Of-Memory issue.

A.3 MLP Fusion Approach

To ensure a fair comparison, we equip conventional forecasting models with an efficient MLP-based fusion module (Huang et al., 2025a), allowing them to incorporate future exogenous information. Given historical endogenous variables $\mathbf{X}^{\text{endo}} \in \mathbb{R}^{N\times T}$ and historical exogenous variables $\mathbf{X}^{\text{exo}} \in \mathbb{R}^{D\times T}$, the forecasting backbone first encodes the historical observations into a latent representation:

$$\mathbf{Z} = \theta_{\text{Model}}\left(\mathbf{X}^{\text{endo}}, \mathbf{X}^{\text{exo}}\right), \tag{19}$$

where $\mathbf{Z} \in \mathbb{R}^{N\times F}$ denotes the learned representation.

The resulting representation is then fused with the future exogenous variables $\mathbf{Y}^{\text{exo}} \in \mathbb{R}^{D\times F}$ through concatenation, followed by an MLP prediction head:

$$\hat{\mathbf{Y}}^{\text{endo}} = \theta_{\text{MLP}}\left(\text{Concat}\left(\mathbf{Z}, \mathbf{Y}^{\text{exo}}\right)\right), \tag{20}$$

where θ_{Model} and θ_{MLP} represent the parameters of the forecasting backbone and the fusion module, respectively.

B Experimental Results

B.1 Hyperparameter Sensitivity

In this section, we investigate the key hyperparameters of NMask, including patch length, model dimension and learning rate. The analysis is conducted on the NP, PJM, DE, and Energy datasets, with a lookback length of 720 and a forecasting horizon of 360 steps. During the analysis, only one hyperparameter is varied at a time, while the others are kept fixed. As shown in Figure 5,

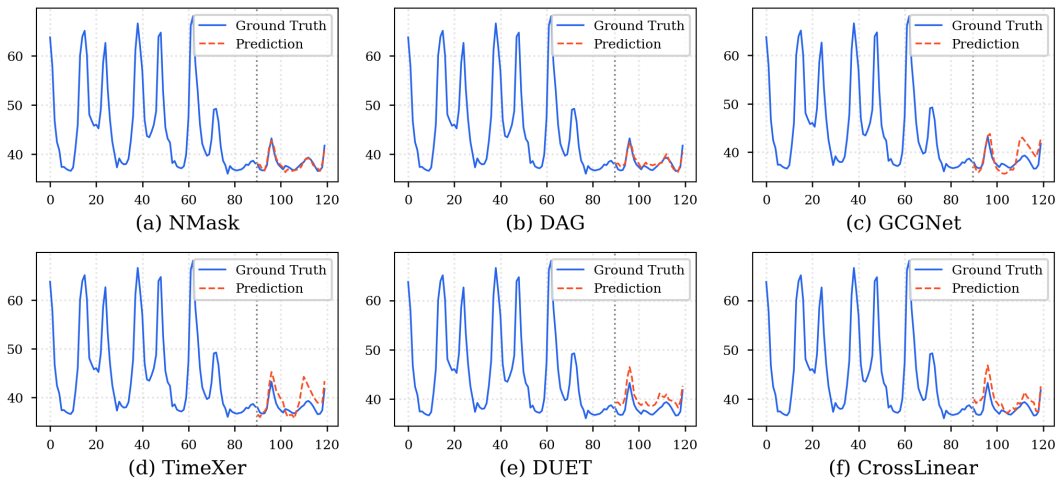


Figure 6: Visualization of prediction results on NP dataset where future exogenous variables are available.

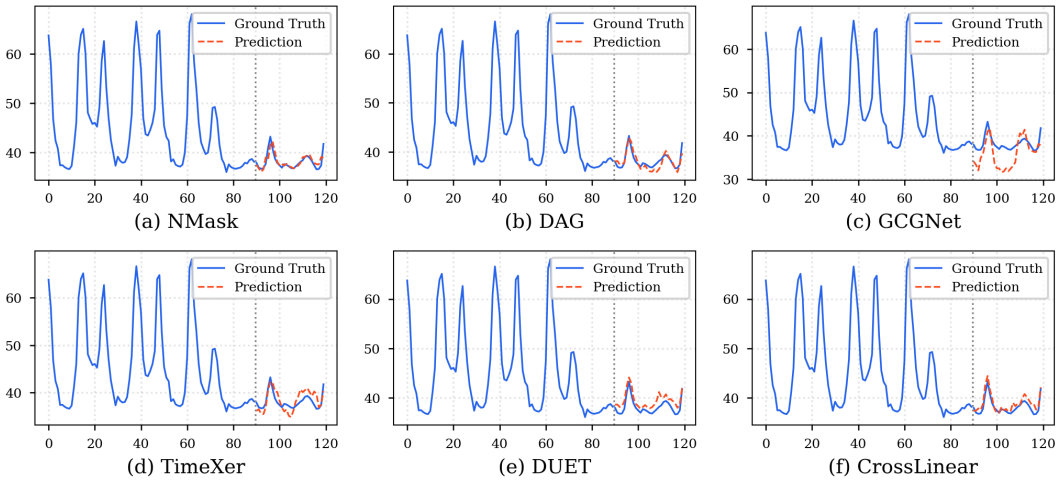


Figure 7: Visualization of prediction results on NP dataset where future exogenous variables are not available.

NMask exhibits stable performance across different configurations, demonstrating its robustness to hyperparameter variations.

For the patch length, NMask achieves the best overall performance with a patch length of around 24 across most datasets. Although the optimal patch length varies among datasets, excessively short or long patches do not lead to significant performance degradation, indicating that NMask can effectively capture temporal patterns at different granularities. For the model dimension, increasing the hidden dimension does not always improve forecasting accuracy. NMask maintains competitive performance with relatively small model dimensions ranging from 64 to 128, suggesting that the proposed masked alignment mechanism and shared predictor can effectively utilize temporal information without relying on large representation capacity.

For the learning rate, NMask achieves relatively better performance with learning rates ranging from 0.001 to 0.005. Extremely small learning rates may slow down the optimization process, while excessively large learning rates can make training less stable. Overall, the results show that NMask is robust to moderate changes in key hyperparameters and achieves reliable forecasting performance across different datasets.

Table 5: Full results on 12 real-world datasets where future exogenous variables are available. The best performance is highlighted in **red**, and the second-best is underlined. Avg represents the average results across the two forecasting horizons.

Models		**NMask**		DAG		GCGNet		TimeXer		TFT		TiDE		DUET		CrossLinear		Amplifier		TimeKAN	
Metrics		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
NP	24	**0.179**	**0.221**	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
	360	**0.437**	**0.419**	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
	Avg	**0.308**	**0.320**	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
PJM	24	0.060	0.147	**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
	360	**0.110**	**0.209**	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
	Avg	**0.085**	**0.178**	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
BE	24	**0.342**	0.230	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
	360	**0.478**	0.326	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
	Avg	**0.410**	**0.278**	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
FR	24	**0.313**	0.175	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
	360	**0.446**	**0.257**	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
	Avg	**0.380**	**0.216**	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
DE	24	**0.249**	**0.313**	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
	360	**0.400**	**0.395**	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
	Avg	**0.325**	**0.354**	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
Energy	24	**0.057**	**0.187**	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
	360	**0.114**	**0.268**	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
	Avg	**0.086**	**0.228**	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
Sdwpfm1	24	0.387	**0.394**	**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
	360	**0.363**	**0.404**	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
	Avg	**0.375**	**0.399**	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
Sdwpfm2	24	0.425	0.425	**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
	360	**0.422**	**0.440**	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
	Avg	**0.424**	**0.432**	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
Sdwpfh1	24	**0.364**	**0.407**	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
	360	**0.366**	**0.415**	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
	Avg	**0.365**	**0.411**	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
Sdwpfh2	24	**0.432**	0.466	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
	360	**0.393**	**0.445**	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
	Avg	**0.413**	**0.456**	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
Colbun	10	0.063	**0.091**	**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
	30	**0.131**	**0.203**	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
	Avg	**0.097**	**0.147**	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
Rapel	10	0.180	0.207	**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
	30	**0.265**	**0.313**	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
	Avg	**0.223**	**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
1st Count		**31**	**29**	5	6	0	0	0	0	0	1	0	0	0	0	0	0	0	0	0	0

B.2 Visualization of Results

We visualize prediction cases on the NP dataset in Figure 6, where future exogenous variables are available, and Figure 7, where they are unavailable. As shown in both figures, NMask consistently achieves more accurate predictions than the baselines, demonstrating the effectiveness of explicitly leveraging the temporal prior between future exogenous variables and future endogenous variables and capturing delayed effects of exogenous variables on endogenous variables through TCEA.

B.3 Full Results

Full results on the 12 multivariate datasets under the time series forecasting setting, where the inputs are $(\mathbf{X}^{\text{endo}}, \mathbf{X}^{\text{exo}}, \mathbf{Y}^{\text{exo}})$, are reported in Table 5. Full results on the multivariate datasets without future exogenous variables, where the inputs are $(\mathbf{X}^{\text{endo}}, \mathbf{X}^{\text{exo}})$, are provided in Table 6.

B.4 Limitations and Future Works

Although NMask effectively leverages known future covariates for time series forecasting, this work mainly focuses on scenarios where future covariates are either available or can be partially estimated

Table 6: Full results on 12 real-world datasets where future exogenous variables are not available. The best performance is highlighted in **red**, and the second-best is underlined. Avg represents the average results across the two forecasting horizons.

Models		**NMask**		DAG		GCGNet		TimeXer		TFT		TiDE		DUET		CrossLinear		Amplifier		TimeKAN	
Metrics		MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
NP	24	**0.217**	**0.245**	0.237	0.259	0.240	0.260	0.270	0.292	0.330	0.320	0.293	0.307	0.260	0.271	0.246	0.282	0.302	0.321	0.336	0.344
	360	0.638	0.487	**0.601**	0.500	0.620	0.502	0.609	**0.474**	0.964	0.655	0.638	0.523	0.625	0.494	0.655	0.505	0.738	0.574	0.632	0.527
	Avg	0.428	**0.366**	**0.419**	0.380	0.430	0.381	0.440	0.383	0.647	0.488	0.466	0.415	0.443	0.383	0.451	0.394	0.520	0.448	0.484	0.436
PJM	24	**0.066**	**0.160**	0.072	0.162	0.072	0.164	0.096	0.191	0.129	0.230	0.104	0.212	0.081	0.175	0.096	0.198	0.093	0.195	0.130	0.236
	360	**0.180**	**0.259**	0.180	0.259	0.190	0.269	0.186	0.267	0.270	0.310	0.211	0.292	0.198	0.272	0.197	0.283	0.210	0.294	0.219	0.304
	Avg	**0.123**	**0.210**	0.126	0.210	0.131	0.216	0.141	0.229	0.200	0.270	0.158	0.252	0.140	0.224	0.147	0.241	0.152	0.245	0.175	0.270
BE	24	**0.350**	**0.236**	0.376	0.244	0.365	0.259	0.392	0.253	0.400	0.267	0.491	0.322	0.375	0.253	0.389	0.253	0.417	0.283	0.424	0.280
	360	0.563	**0.343**	**0.547**	0.350	0.553	0.354	0.561	0.349	0.726	0.440	0.602	0.375	0.571	0.357	0.564	0.346	0.588	0.367	0.552	0.350
	Avg	**0.457**	**0.290**	0.462	0.297	0.459	0.306	0.477	0.301	0.563	0.354	0.547	0.349	0.473	0.305	0.477	0.300	0.503	0.325	0.488	0.315
FR	24	**0.339**	**0.186**	0.345	0.190	0.352	0.199	0.366	0.195	0.445	0.231	0.411	0.253	0.356	0.207	0.397	0.204	0.429	0.250	0.448	0.250
	360	0.523	**0.292**	0.526	0.307	**0.519**	0.293	0.542	0.298	0.624	0.345	0.574	0.325	0.578	0.317	0.555	0.310	0.559	0.322	0.534	0.303
	Avg	**0.431**	**0.239**	0.435	0.249	0.435	0.246	0.454	0.247	0.535	0.288	0.493	0.289	0.467	0.262	0.476	0.257	0.494	0.286	0.491	0.277
DE	24	0.418	**0.397**	0.422	0.411	**0.413**	0.402	0.501	0.445	0.576	0.452	0.464	0.433	0.423	0.403	0.438	0.423	0.493	0.455	0.504	0.462
	360	**0.769**	**0.557**	0.784	0.563	0.815	0.600	0.818	0.568	0.792	0.578	0.821	0.604	0.895	0.623	0.832	0.592	0.930	0.642	0.834	0.590
	Avg	**0.593**	**0.477**	0.603	0.487	0.614	0.501	0.660	0.507	0.684	0.515	0.643	0.519	0.659	0.513	0.635	0.508	0.712	0.549	0.669	0.526
Energy	24	0.120	0.269	0.112	0.257	0.112	0.259	0.138	0.292	0.350	0.462	0.116	0.265	0.109	0.254	**0.102**	**0.246**	0.108	0.254	0.109	0.254
	360	0.197	0.347	0.189	0.342	0.195	0.351	0.206	0.360	0.399	0.497	0.209	0.358	0.211	0.362	0.299	0.425	0.251	0.400	**0.186**	**0.338**
	Avg	0.159	0.308	0.151	0.300	0.154	0.305	0.172	0.326	0.375	0.480	0.163	0.312	0.160	0.308	0.201	0.336	0.180	0.327	**0.148**	**0.296**
Sdwpfm1	24	**0.531**	**0.477**	0.532	0.512	0.537	0.509	0.558	0.533	0.596	0.531	0.560	0.541	0.550	0.496	0.549	0.553	0.579	0.557	0.598	0.579
	360	0.879	**0.662**	0.845	0.695	0.885	0.682	**0.845**	0.684	1.372	0.925	0.863	0.725	0.898	0.673	1.070	0.834	1.106	0.814	0.853	0.764
	Avg	0.705	**0.570**	**0.689**	0.604	0.711	0.596	0.702	0.609	0.984	0.728	0.712	0.633	0.724	0.585	0.810	0.694	0.843	0.686	0.726	0.672
Sdwpfm2	24	**0.598**	**0.526**	0.609	0.548	0.630	0.547	0.627	0.570	0.744	0.655	0.636	0.582	0.610	0.537	0.635	0.586	0.669	0.600	0.684	0.621
	360	**0.958**	**0.704**	0.962	0.766	0.983	0.739	0.978	0.736	1.340	0.877	0.995	0.783	1.026	0.744	0.965	0.788	1.215	0.863	0.987	0.780
	Avg	**0.778**	**0.615**	0.786	0.657	0.807	0.643	0.803	0.653	1.042	0.766	0.816	0.683	0.818	0.641	0.800	0.687	0.942	0.732	0.836	0.701
Sdwpfh1	24	**0.645**	**0.574**	0.647	0.591	0.651	0.578	0.651	0.585	0.754	0.670	0.713	0.640	0.673	0.575	0.687	0.652	1.111	0.816	0.755	0.679
	360	0.888	**0.670**	**0.820**	0.694	0.830	0.677	0.841	0.700	0.831	0.724	0.896	0.755	0.886	0.712	0.849	0.752	1.089	0.823	0.853	0.762
	Avg	0.767	**0.622**	**0.734**	0.643	0.741	0.628	0.746	0.643	0.793	0.697	0.805	0.698	0.780	0.644	0.768	0.702	1.100	0.820	0.804	0.721
Sdwpfh2	24	0.748	**0.611**	**0.739**	0.635	0.798	0.645	0.820	0.677	0.811	0.714	0.784	0.674	0.974	0.692	0.810	0.718	0.834	0.691	0.892	0.740
	360	**0.962**	**0.726**	1.000	0.755	0.974	0.757	0.962	0.761	1.041	0.777	1.056	0.829	1.039	0.737	1.102	0.830	1.126	0.845	0.990	0.825
	Avg	**0.855**	**0.669**	0.870	0.695	0.886	0.701	0.891	0.719	0.926	0.746	0.920	0.752	1.007	0.715	0.956	0.774	0.980	0.768	0.941	0.783
Colbun	10	**0.062**	**0.089**	0.072	0.102	0.070	0.094	0.084	0.146	0.449	0.279	0.094	0.140	0.071	0.103	0.068	0.101	0.077	0.120	0.091	0.129
	30	**0.159**	0.211	0.161	**0.208**	0.168	0.236	0.180	0.290	0.663	0.494	0.280	0.343	0.221	0.290	0.190	0.289	0.227	0.301	0.311	0.377
	Avg	**0.110**	**0.150**	0.117	0.155	0.119	0.165	0.132	0.218	0.556	0.387	0.187	0.242	0.146	0.197	0.129	0.195	0.152	0.211	0.201	0.253
Rapel	10	**0.195**	0.221	0.195	**0.215**	0.217	0.221	0.301	0.308	0.238	0.274	0.239	0.270	0.222	0.221	0.201	0.224	0.240	0.259	0.226	0.238
	30	0.321	**0.327**	0.327	0.347	0.313	0.346	0.387	0.416	0.377	0.394	0.409	0.444	0.387	0.386	**0.280**	0.374	0.433	0.436	0.458	0.437
	Avg	0.258	**0.274**	0.261	0.281	0.265	0.284	0.344	0.362	0.308	0.334	0.324	0.357	0.305	0.304	**0.241**	0.299	0.337	0.348	0.342	0.338
1st Count		**21**	**30**	7	2	2	0	1	1	0	0	0	0	0	0	3	1	0	0	2	2

with reasonable accuracy. In practice, future covariates may contain prediction errors or uncertainty, which could affect their reliability. Future work will explore uncertainty-aware modeling strategies to better utilize imperfect future covariates. In addition, while NMask captures delayed effects through TCEA, future work could further investigate more adaptive strategies for modeling complex dependencies among exogenous variables.