IEEE Access
Multidisciplinary | Rapid Review | Open Access Journal

Date of publication xxxx 00, 0000, date of current version xxxx 00, 0000.

Digital Object Identifier 10.1109/ACCESS.2017.Doi Number

Distributed Lag Transformer based on Time-Variable-Aware Learning for Explainable Multivariate Time Series Forecasting

YOUNGHWI KIM[1], DOHEE KIM[2](Member, IEEE), JOONGROCK KIM[2], SUNGHYUN SIM[2*] (Member, IEEE)

[1]Graduate School of AI Convergence Engineering, Changwon National University, 20 Changwondaehak-ro Uichang Gu, 51140 Changwon, South Korea,
[2]School of AI Convergence Engineering, Changwon National University, 20 Changwondaehak-ro Uichang Gu, 51140 Changwon, South Korea

Younghwi Kim and Dohee Kim are co-first authors.

*Corresponding author: Sunghyun Sim (ssh@changwon.ac.kr)

This work was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (No. RS-2023-00218913) and was supported by the National Research Foundation of Korea (NRF) grant funded by the Korean government (MSIT) (No.RS-2025-25396743).

ABSTRACT Time series data is a key element of big data analytics, commonly found in domains such as finance, healthcare, climate forecasting, and transportation. In large-scale real-world settings, such data is often high-dimensional and multivariate, requiring advanced forecasting methods that are both accurate and interpretable. Although Transformer-based models have achieved strong performance in multivariate time series forecasting (MTSF), their lack of explainability limits their use in critical applications. To address this limitation, we propose the Distributed Lag Transformer (DLFormer), a novel Transformer architecture for explainable and scalable MTSF. DLFormer integrates distributed lag embedding and time-variable-aware learning (TVAL) to structurally model both local and global temporal dependencies and explicitly capture the influence of past variables on future outcomes. Experiments on ten benchmark and real-world datasets demonstrate that DLFormer achieves competitive predictive accuracy, yielding MSE reductions of 1.41% on average across all settings and 2.66% in short-term forecasting relative to three top-performing baselines, including iTransformer, NLinear, and TFT. Beyond predictive accuracy, we evaluate the quality of explainability through global attention-based analysis, perturbation-based faithfulness evaluation, and domain-oriented case studies. These analyses show that DLFormer provides more robust, faithful, and valid explanations of temporal-variable dependencies than competing models. Overall, these results suggest that DLFormer effectively bridges the gap between predictive performance and explainability. The source code is publicly available at: http://github.com/kYounghwi/DLFormer.

INDEX TERMS Distributed Lag Transformer, Distributed Lag Embedding, Explainable Multivariate Time Series Forecasting, Time-Variable-Aware Learning

I. INTRODUCTION

Time series, which consists of sequential observations recorded over time, serves as a fundamental data type in big data analytics across a wide range of domains including finance, healthcare, climate forecasting, and traffic management [1]–[4]. In real-world big data environments, time series data is typically multivariate and high-dimensional, encompassing numerous interdependent variables that evolve simultaneously over time [5]–[7]. This complexity gives rise to the critical problem of multivariate time series forecasting (MTSF) in big data systems, which demands the modeling of both short- and long-range temporal dependencies to generate accurate and interpretable predictions at scale [8]–[10].

Traditional approaches to the MTSF problem include statistical models (e.g., Autoregressive Integrated Moving Average (ARIMA) and Vector Autoregression (VAR)) and deep learning-based models (e.g., Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM), and Transformers) that have been extensively studied. While statistical models provide explainable structures, they often struggle to capture complex temporal dependencies, especially in high-dimensional data [11]. In contrast, deep

learning models have demonstrated remarkable predictive performance but lack explainability, hindering the interpretability of predictions [12]. This limitation is critical in domains where transparency and trust in predictions are essential, such as finance and healthcare [13].

Although deep learning-based approaches such as Transformer models have shown continuous advancement, several critical limitations still persist.

- ***Lack of explainability for temporal relationships among variable***s: Transformer-based models leverage attention mechanisms to capture dependencies among variables, yet they do not explicitly model how historical values of individual variables impact future predictions. Consequently, visualizing attention weights is insufficient for quantifying the temporal impact of individual variables, complicating the interpretation of the underlying forecasting dynamics.
- ***Limitations of explainable artificial intelligence (XAI) methods in time series forecasting***: Recent advancements in XAI have primarily targeted non-temporal data domains such as images and text. Conventional XAI techniques, such as Shapley Additive Explanations (SHAP) and Local Interpretable Model-agnostic Explanations (LIME), struggle to incorporate the temporal structure of multivariate time series data, making them inadequate for explaining the lagged effects of variables on future predictions [12].
- ***Trade-off between predictive performance and explainability***: Existing research often prioritized improving predictive accuracy in Transformer-based forecasting models while overlooking rigorous evaluations of explainability. High explainability models commonly demonstrate reduced predictive performance, and effective methods to balance both aspects remain lacking [14].

To address these challenges, we propose a novel approach that enhances the explainability of the MTSF problem while maintaining high predictive performance. Specifically, we introduce time-variable-aware learning (TVAL), which explicitly models the contributions of individual variables' past values (temporal lags) to future predictions.

To enable TVAL, we developed the Distributed Lag Transformer (DLFormer), which integrates the well-established distributed lag mechanism [15] from econometrics into Transformer-based time series models. DLFormer decomposes each variable's temporal information, creating a structured representation of local (within-variable) and global (across-variable) temporal dependencies. The core component, distributed lag embedding (DL Embedding), enables the model to effectively learn and quantify the relative influence of past observations on its future predictions.

The main novelty of this study lies in reformulating the Transformer architecture from a time-variable-aware perspective for explainable MTSF, where variable-specific temporal lags are explicitly represented as token-level modeling units. This design enables DLFormer to jointly capture local and global temporal dependencies while providing structured explanations from both temporal and variable perspectives. The main contributions of this study are summarized as follows:

- **(C1)** We introduce TVAL, a new learning paradigm for MTSF that explicitly captures variable-specific temporal dependencies, enhancing interpretability and predictive performance.
- **(C2)** We propose DLFormer, the first Transformer-based model that structurally integrates the distributed lag mechanism to model the impact of past observations on future predictions.
- **(C3)** We develop DL Embedding, a novel embedding strategy that encodes local and global temporal relationships in a structured and interpretable representation.
- **(C4)** Extensive experiments on ten benchmark and real-world datasets demonstrate that DLFormer achieves competitive predictive accuracy while significantly enhancing model explainability from temporal and variable perspectives.

The remainder of this paper is organized as follows. Section II reviews existing XAI methods and explainable time-series forecasting approaches. Section III introduces the proposed DLFormer architecture, including DL Embedding, the DL Encoder–Decoder structure, and the TVA attention map. Section IV describes the experimental details and presents the main forecasting and explainability results. Section V provides further analyses, including ablation studies, scalability analysis, and qualitative interpretation of the learned attention patterns. Section VI discusses the methodological implications, limitations, and extensibility of DLFormer based on the preceding experimental results. Finally, Section VII concludes the paper by summarizing the main findings and outlining future research directions.

II. RELATED WORK

This section reviews XAI methods and discusses studies that apply these approaches to multivariate time-series data.

A. Explainable Artificial Intelligence

With the increasing adoption of machine learning and deep learning models across various industries, the demand for XAI has grown significantly [16]. The "black-box problem" in deep learning [17], where models produce highly accurate results but lack transparency in decision-making, has spurred extensive research in XAI [18]. The core objective of XAI is to provide understandable explanations for AI decisions, enabling stakeholders to trust and validate model outputs [12]. XAI methods can be categorized across three independent dimensions: the scope of explanation, timing of explanation, and methodological dependency [11], [12].

TABLE I
EXPLAINABLE TIME-SERIES FORECASTING MODELS

Type	Model	Explainer	Input Series (Multi / Uni)	Explainable Dimension (Multi / Uni)	Evaluation Scope (Global / Local)	Evaluation (Qualitative / Quantitative)
Post-hoc	XCM (2021) [30]	Grad-CAM	M	M	Local	Quali / Quant
	TimeSHAP (2021) [39]	SHAP	M	M	Global / Local	Quali
	ForecastCF (2023) [41]	Counterfactual	U	U	Local	Quali / Quant
	MTEX (2024) [27]	Integrated-Gradient	M	M	Local	Quali / Quant
	TIMING (2025) [38]	Integrated-Gradient	M	U	Local	Quali / Quant
	ChronoSHAP (2026) [40]	SHAP	M	U	Global	Quali
Ante-hoc	IMV-LSTM (2019) [37]	Attention	M	M	Global	Quali / Quant
	TFT (2021) [28]	Attention / VSN	M	M	Global	Qual
	TACN (2020) [29]	Attention	M	U	Local	Qual
	Autoformer (2021) [35]	Auto-Correlation	M	U	Local	Quali
	PatchTST (2023) [34]	Attention	M	U	Local	Quali
	TimesNet (2023) [36]	Attention	M	U	Local	Quali
	NLinear (2023) [33]	Linear	U	U	Global	Qual
	iTransformer (2024) [32]	Attention	M	U	Global	Qual
	Ours (DLFormer)	**Attention**	**M**	**M**	**Global / Local**	**Quali / Quant**

First, regarding the scope of explanation, XAI approaches can be divided into global and local explanations. Global explanations aim to understand the model's overall decision-making process, deriving general rules and patterns applicable to the entire dataset. Analyzing feature importance and decision pathways allows researchers and practitioners to assess whether the model behaves logically and consistently [19]. In contrast, local explanations aim to explain individual predictions by analyzing how specific inputs led to specific outputs. This is crucial in high-stakes domains such as medical diagnostics and fraud detection, where understanding the rationale behind individual decisions is critical [18], [20]. Modern XAI research seeks to balance both explanations, but the complexity of deep learning models presents challenges, resulting in many methods focusing on one type over the other [21].

Second, based on the timing of explanation, XAI methods are categorized into post-hoc and ante-hoc approaches. Post-hoc methods provide explanations after model training, making them independent of the underlying architecture.

These methods are widely used because they introduce interpretability to existing models without altering their internal structure [12]. In contrast, ante-hoc methods are designed with transparency as a core principle. Often referred to as White-box models, these methods ensure that decision-making processes are inherently interpretable without additional tools [18].

Third, from the perspective of methodological dependency, XAI methods can be classified into model-agnostic and model-specific techniques. Model-agnostic methods, such as SHAP [22], LIME [23], and Gradient-based approaches, can be applied universally across different machine learning models. In contrast, model-specific methods are tailored for specific model architectures, particularly deep learning models such as Gradient-weighted Class Activation Mapping (Grad-CAM) [24] and Grad-CAM++ [25]. These methods leverage internal model information to generate more precise and architecture-aware explanations but lack generalizability across various datasets and tasks [26].

B. Explainable Method for Time-series forecasting

Time-series forecasting involves nonlinear temporal dependencies and delayed interactions among multiple variables, complicating the development and evaluation of explainable methods. Existing XAI methods, such as LIME and SHAP, were originally designed for models handling cross-sectional data, limiting their direct application to time-series tasks [12], [27]. Furthermore, forecasting requires more intricate dynamics to be explained compared to classification problems, as it models long-term dependencies and interactions over time [28]. Consequently, recent research has focused on adapting existing XAI methods to better suit time-series forecasting or specifically designing new ante-hoc approaches [27]-[30].

Explainable time-series forecasting methods can be classified based on their emphasis on variable dimension (variable importance (VI)) or temporal dimension (temporal importance (TI)) [31]. Accordingly, explainability methods can be divided into uni-dimensional explainable (UDE) models, providing explanations for one dimension (feature or time), and multi-dimensional explainable (MDE) models, which consider both dimensions simultaneously. UDE methods primarily analyze specific components of a model, often emphasizing predictive performance over explainability.

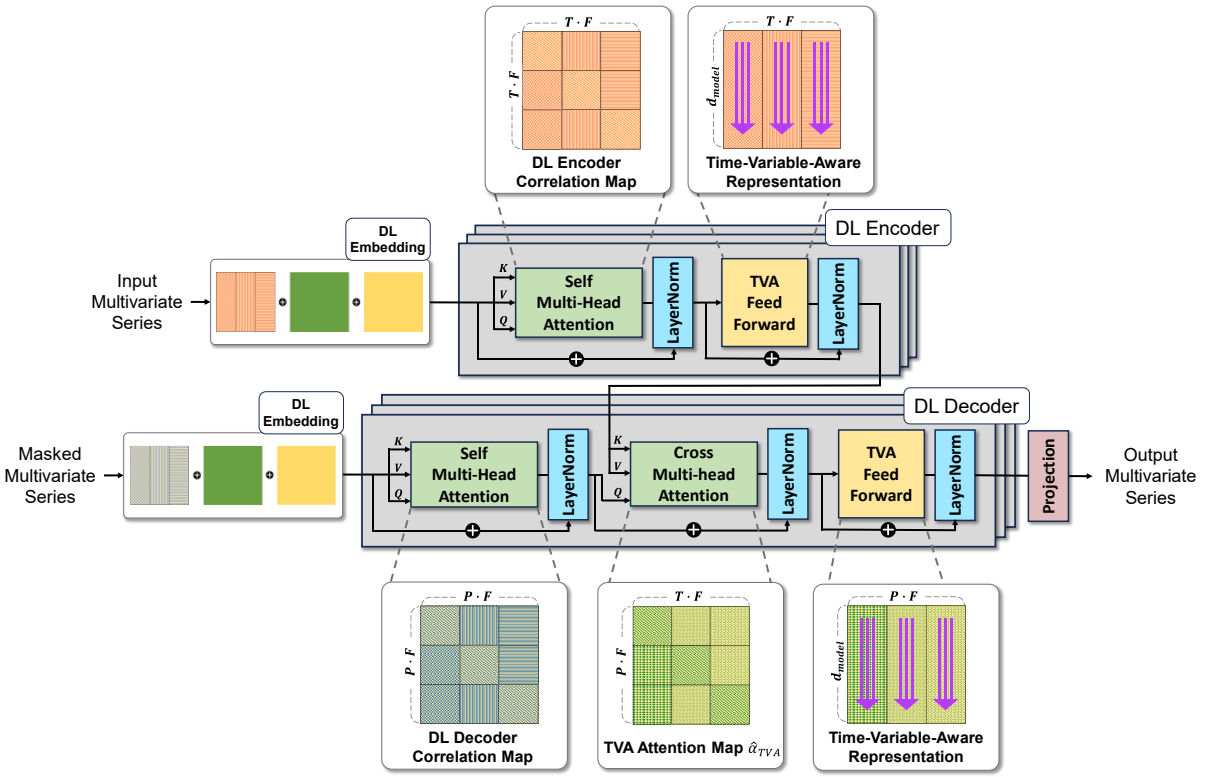

FIGURE 1. Proposed DLFormer network structure.

In contrast, MDE methods aim to provide a more comprehensive explanation by modeling relationships between time and variables; however, they entail increased computational complexity, which can impact predictive performance.

Notable UDE methods include Temporal Attention Convolutional Neural Networks (TACN) [29], iTransformer [32], and NLinear [33]. TACN applies attention mechanisms to time tokens to highlight TI, integrating this with Temporal Convolutional Networks to form a structured model. iTransformer reverses the standard transformer input structure, treating time and variable dimensions interchangeably, thereby capturing variable interactions using attention mechanisms. NLinear employs explicit autoregressive linear layers for each time series, effectively modeling autoregressive temporal patterns. Beyond these approaches, there have also been studies that introduce inductive biases specialized for long-term time-series. For instance, PatchTST [34] partitions time series into patch-level temporal tokens and feeds them into a Transformer, thereby enabling effective modeling of long-range dependencies. Autoformer [35] leverages an auto-correlation mechanism to capture periodic patterns in complex temporal dynamics. Likewise, TimesNet [36] reconstructs one-dimensional time series from a two-dimensional variation perspective, allowing the model to effectively learn multi-periodic patterns. However, these models were primarily designed to improve predictive accuracy rather than to provide explicitly structured explanations over both temporal and variable dimensions. As a result, systematic quantitative and qualitative evaluations of explainability have remained relatively underexplored in this line of research.

In contrast, MDE methods integrate variable and temporal dependencies within a unified interpretability framework. Representative models include Interpretable Multi-Variable (IMV)-LSTM [37] and Temporal Fusion Transformer (TFT) [28]. IMV-LSTM extends conventional LSTMs by learning variable-wise hidden states and employing a mixture attention mechanism to jointly capture variable importance and temporal importance during prediction. In contrast, TFT combines attention-based estimation of temporal importance with an internal variable selection network (VSN), thereby enabling separate evaluation of VI and TI. Although MDE methods generally provide more comprehensive explanations than UDE methods, they may be disadvantaged in predictive performance because they must preserve interpretability while modeling complex temporal dynamics.

In addition, post-hoc explainability methods for time-series forecasting have also been actively studied. Representative gradient-based approaches include XCM [30] and MTEX [27], which adapt Grad-CAM and Integrated Gradients, respectively, to CNN-based architectures optimized for time-series tasks. More recently, temporality-aware attribution methods such as TIMING [38] have also been proposed by extending Integrated Gradients to better reflect temporal dependencies in time series. SHAP-based approaches include

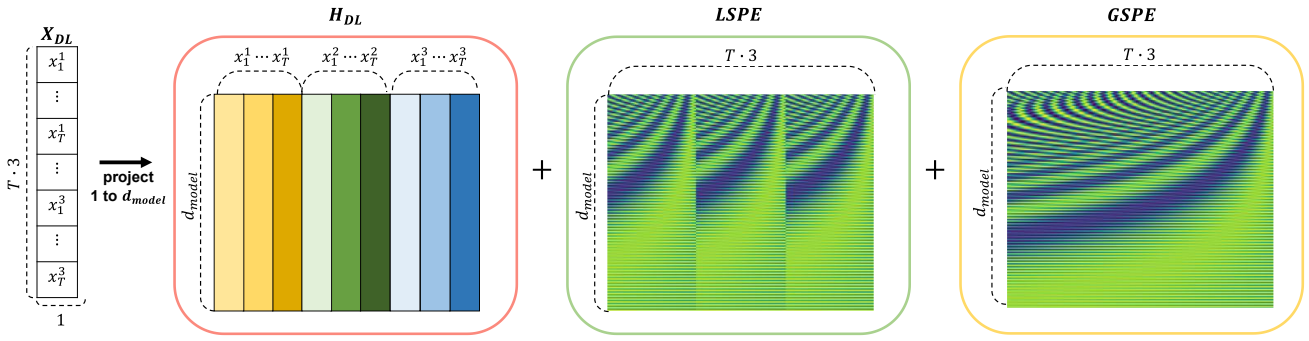

FIGURE 2. Detailed visualization of Distributed Lag Embedding operation.

TimeSHAP [39], a model-agnostic explainer for recurrent models, and ChronoSHAP [40], which analyzes global temporal dependencies in multivariate multi-step forecasting. In addition, counterfactual-based approaches such as ForecastCF [41] explain forecasting results by identifying input perturbations that induce desired changes in the predicted future trajectory. However, although these post-hoc methods can provide useful explanations, unlike ante-hoc approaches that offer structurally integrated explanations within the model, they
analyze an already trained predictor from outside the forecasting process, and thus the consistency and form of the resulting explanations may vary depending on the explainer and the underlying predictive model.

Taken together, these discussions suggest that, in time-series forecasting, MDE explanations that jointly consider temporal and variable dimensions are important. However, existing ante-hoc approaches may suffer from degraded predictive performance as they attempt to preserve interpretability, while post-hoc approaches may be limited in directly reflecting the structural characteristics of the model itself. This perspective constitutes the central motivation of the present study, which aims to achieve strong predictive performance and structured interpretability simultaneously within a unified modeling framework. Accordingly, we propose DLFormer, an ante-hoc forecasting model with MDE explainability.

TABLE II
SUMMARY OF NOTATIONS

Symbols	Description
S, F	Number of time series samples and variables, respectively
T, P	Input sequence length and forecasting horizon
N, M, d	Number of encoder layer, decoder layer, and embedding dimension
$X_{1:T} \in \mathbb{R}^{S\times T\times F}$	Input multivariate time series with T sequence
$X_{T:T+P} \in \mathbb{R}^{S\times P\times F}$	Target sequence representing the P future observations
$X_{DL} \in \mathbb{R}^{S\times T\cdot F\times 1}$	Input to the distributed lag embedding
$H_{DL} \in \mathbb{R}^{S\times T\cdot F\times d_{model}}$	Embedded representation of X_{DL}
$z^{enc} \in \mathbb{R}^{S\times T\cdot F\times d_{model}}$	Encoder output
$z^{dec} \in \mathbb{R}^{S\times T\cdot F\times d_{model}}$	Decoder output
$\hat{\alpha}$	Explanation generated by DLFormer for time-series
$\ell, \varphi, \nabla_{\varphi}$	Loss function, model parameters, and gradient of φ, respectively

III. DISTRIBUTED LAG TRANSFORMER

This section introduces the novel explainable MTSF model DLFormer and provides a detailed explanation of the TVAL method within it. Table II summarizes the notations used, while Fig. 1 illustrates the overall architecture of the proposed DLFormer. DLFormer comprises three main components: (1) DL Embedding, (2) Distributed Lag Encoder (DL Encoder), and (3) Distributed Lag Decoder (DL Decoder).

A. Distributed Lag Embedding

The DL Embedding module transforms a multivariate time-series input into a flattened time-variable token sequence. Given F variables over T time steps, the input is first rearranged by concatenating the univariate sequences, as defined in (1):

$$X_{DL} = \{x_1^1, \cdots, x_T^1, \cdots, x_1^F, \cdots, x_T^F\}. \tag{1}$$

A linear projection then maps each scalar observation in the flattened sequence into a d_{model}-dimensional embedding, i.e., $X_{DL} \in \mathbb{R}^{S\times T\cdot F\times 1} \rightarrow H_{DL} \in \mathbb{R}^{S\times T\cdot F\times d_{model}}$, where d_{model} denotes the embedding dimension. Through this flattening process, each token corresponds to a specific variable at a specific temporal lag, thereby serving as the basic modeling and explanation unit of DLFormer.

Subsequently, to preserve the identity of these flattened tokens, DLFormer introduces two positional encodings: local sequence position embedding (LSPE) and global sequence position embedding (GSPE). Both are derived from the sinusoidal sequence position embedding (SPE) used in the original Transformer architecture [42], defined as follows:

$$SPE_{t,d} = \begin{cases} \sin(t/C^{d/d_{model}}), if\ d\ mod\ 2 = 0 \\ \cos(t/C^{d/d_{model}}), \quad otherwise \end{cases}. \tag{2}$$

Here, $t \in \{1, \dots, T\}$ denotes the position index, $d \in \{0, \dots, d_{model}\}$ denotes the embedding-dimension index, and C is a constant that determines the periodicity. LSPE is reset for each variable and therefore encodes the local temporal order within each univariate sequence. It is constructed by concatenating F variable-wise SPEs, as follows:

$$LSPE_{t,d} = Concat(SPE_{t,d}^{1}, \cdots, SPE_{t,d}^{F}). \quad (3)$$

In contrast, GSPE is assigned over the entire flattened sequence by defining the position range as $p \in \{1, \dots, T \cdot F\}$. Therefore, GSPE encodes the global time-variable position of each flattened token. The final DL Embedding is obtained by summing the projected input, LSPE, and GSPE:

$$DLE = H_{DL} + GSPE_{p,d} + LSPE_{p,d}. \quad (4)$$

Here, p denotes the flattened token index. Through this dual positional design, each token retains both its within-variable lag identity and global variable-lag identity, enabling the model to distinguish temporal continuity within a variable from lagged interactions across variables.

Fig. 2 illustrates the process of applying DL Embedding to a multivariate time series dataset consisting of three univariate time series. First, the multivariate time series data, composed of three univariate time series, is rearranged into a single-dimensional sequence to construct X_{DL}. Next, a linear transformation with dimensionality d_{model} is applied to X_{DL} obtain H_{DL} which has dimensions $d_{model} \times T \cdot 3$. LSPE is obtained by combining three SPEs, each with a dimensionality of $d_{model} \times T$. In contrast, GSPE is constructed as a single SPE with dimensionality $d_{model} \times T \cdot 3$. Finally, the embedded vector is obtained by summing H_{DL}, LSPE, and GSPE.

In Transformers, positional information plays an important role in distinguishing the order and identity of tokens within a sequence. Based on this principle, DLFormer introduces a dual positional inductive bias into the flattened time-variable sequence through LSPE and GSPE. LSPE preserves the temporal order within each variable, whereas GSPE encodes the global time-variable position of each token in the flattened sequence. As a result, each token can retain both its local lag identity and global time-variable identity, helping the model distinguish within-variable temporal continuity from cross-variable lagged interactions. The importance of this design is further supported by the ablation results for LSPE and GSPE.

B. Distributed Lag Encoder and Distributed Lag Decoder

The DL Encoder and DL Decoder combine a standard Transformer with DL Embedding. The model consists of N encoders and M decoders, each learning through the following attention mechanism:

$$A(Q, K) = Softmax\left(QK^{\mathrm{T}}/\sqrt{d_{attn}}\right), \quad (5)$$

$$Attention(Q, K, V) = A(Q, K)V, \quad (6)$$

$$H_i = Attention(QW_i^Q, KW_i^K, VW_i^V). \quad (7)$$

In (7), inputs Q, K, and V denote the DL Embedding vectors, each with an embedding dimension of d_{model}. The parameter matrices W_i^Q, W_i^K, W_i^V correspond to the i^{th} attention head, applied to Q, K, and V, respectively. The output H_i represents the result from the i^{th} attention head. To learn from multiple perspectives, multi-head attention (MHA) performs the operation I times, following linear transformation of the input embedding dimension to d_{attn}, as follows:

$$MHA(Q, K, V) = Concat(H_1, \cdots, H_I)W^M. \quad (8)$$

The DL Encoder utilizes the DL Embedding vector as the input to Self-MHA for learning. This process is formulated as follows:

$$z_{n,1}^{enc} = MHA\left(z_{n-1,2}^{enc}, z_{n-1,2}^{enc}, z_{n-1,2}^{enc}\right) + z_{n-1,2}^{enc}, \quad (9)$$

$$z_{n,2}^{enc} = TVAFF\left(z_{n,1}^{enc}\right) + z_{n,1}^{enc}. \quad (10)$$

Here, $z_{n,2}^{enc}\,(1 \le n \le N)$ represents the output of the nth encoder, which serves as the input for the n+1th encoder. For the first encoder, the input $z_{0,2}^{enc}$ is set to DLE, as in (4).

The time-variable-aware (TVA) feed forward consists of two linear transformations with an activation function in between. TVA feed forward enables the learning of intrinsic characteristics by embedding the representations of each variable at each time step through dense nonlinear connections in multivariate time series. The DL Decoder is structured as follows:

$$z_{m,1}^{dec} = MHA\left(z_{m-1,3}^{dec}, z_{m-1,3}^{dec}, z_{m-1,3}^{dec}\right) + z_{m-1,3}^{dec}, \quad (11)$$

$$z_{m,2}^{dec} = MHA\left(z_{m,1}^{dec}, z_{N,2}^{enc}, z_{N,2}^{enc}\right) + z_{m,1}^{dec}, \quad (12)$$

$$z_{m,3}^{dec} = TVAFF\left(z_{m,2}^{dec}\right) + z_{m,2}^{dec}. \quad (13)$$

Here, $z_{m,3}^{dec}\,(1 \le m \le M)$ represents the output of the m^{th} decoder, which serves as the input for the subsequent decoder. For the DL Embedding input of the initial decoder $z_{0,3}^{dec}$, we use a zero-masked vector of future time steps $\boldsymbol{X}_{T:T+P}^{0}$. As formulated in (11), the decoder's attention mechanism applies Self-MHA to $z_{m-1,3}^{dec}$; subsequently, it applies Cross-MHA as in (12), where the output from the previous Self-MHA serves as Q, while the final encoder output serves as K and V. The decoder generates the final output by embedding the variable-wise temporal representations through TVA feed forward as in (13).

The above steps define the encoding and decoding process of variable-wise temporal representation, where DL embedding, structured attention, and TVA feed forward are

combined within DLFormer, forming the TVAL method. This approach offers stronger representational expressiveness than conventional Transformer-based models that embed variables or time tokens. Finally, future predictions are derived through a single linear transformation, as follows:

$$\hat{X}_{T:T+P} = Projection\left(z_{M,3}^{dec}\right). \quad (14)$$

Because DL Embedding flattens a multivariate input into a sequence of length $T \times F$, it increases the attention sequence length compared with Transformer variants that attend only over T time tokens or F variable tokens. As a result, the self-attention operation has a computational complexity of $O((TF)^2 d_{model})$ and requires $O((TF)^2)$ memory to store the attention map. Although this design increases computational cost, it is a core component of TVAL that enables DLFormer to explicitly model variable-specific lag dependencies. This performance-cost trade-off is further analyzed in the ablation study and discussed in Section VI.

Algorithm 1 presents the pseudocode for training the DLFormer.

Algorithm 1 Training procedure

Input: Time series data $X_{1,T}, X_{T:T+P}^{0}$
$z_{0,2}^{enc} \leftarrow DLE$ applied to $X_{1,T}$
$z_{0,3}^{dec} \leftarrow DLE$ applied to $X_{T:T+P}^{0}$
for $epoch = 1,\ldots, MaxEpoch$ **do**
  for $n = 1,\ldots, N$ **do** ▷ DL Encoder
    $z_{n,1}^{enc} \leftarrow MHA\left(z_{n-1,2}^{enc}, z_{n-1,2}^{enc}, z_{n-1,2}^{enc}\right) + z_{n-1,2}^{enc}$
    $z_{n,2}^{enc} \leftarrow TVAFF\left(z_{n,1}^{enc}\right) + z_{n,1}^{enc}$
  end for
  for $m = 1,\ldots, M$ **do** ▷ DL Decoder
    $z_{m,1}^{dec} \leftarrow MHA\left(z_{m-1,3}^{dec}, z_{m-1,3}^{dec}, z_{m-1,3}^{dec}\right) + z_{m-1,3}^{dec}$
    $z_{m,2}^{dec} \leftarrow MHA\left(z_{m,1}^{dec}, z_{N,2}^{enc}, z_{N,2}^{enc}\right) + z_{m,1}^{dec}$
    $z_{m,3}^{dec} \leftarrow TVAFF\left(z_{m,2}^{dec}\right) + z_{m,2}^{dec}$
  end for
  $\hat{X}_{T:T+P} = Projection\left(z_{M,3}^{dec}\right)$
  $Loss_{\varphi} \leftarrow \ell\left(X_{T:T+P}, \hat{X}_{T:T+P}\right)$
  $\varphi \leftarrow \varphi - \eta \nabla_{\varphi} Loss_{\varphi}$ ▷ Weight update
end for

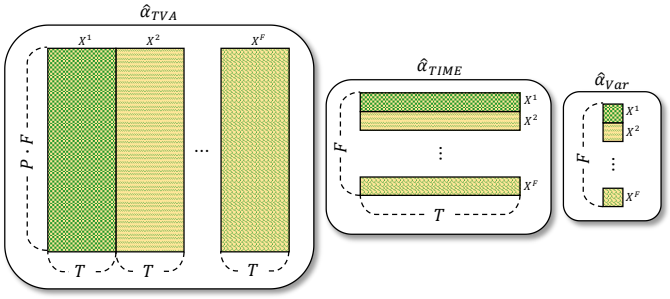

FIGURE 3. Visualization example of the time-variable-aware

C. Time-Variable-Aware Attention Map

In the DLFormer decoder, Cross-MHA is applied as formulated in (12), where the decoder representation serves as Q, and the final encoder representation serves as K and V. This mechanism allows each future query to attend to the flattened variable-lag tokens encoded from the input sequence. Therefore, the resulting cross-attention weights can be interpreted as a model-internal saliency map that indicates the relative predictive importance of input variables across temporal lags. We refer to this map as the TVA attention map, $\hat{\alpha}_{TVA}$, which is computed from the final decoder layer by averaging the attention maps over all heads, as in (15):

$$\hat{\alpha}_{TVA} = \frac{1}{I}\sum_i A_i\left(z_{M,1}^{dec}, z_{N,2}^{enc}\right), \quad (15)$$

$$\hat{\alpha}_{Time} = reshape\left(\frac{1}{PF}\sum_{pf} \hat{\alpha}_{TVA}\right), \quad (16)$$

$$\hat{\alpha}_{Var} = \frac{1}{T}\sum_t \hat{\alpha}_{Time}. \quad (17)$$

Because $\hat{\alpha}_{TVA}$ is defined over both forecasting steps and flattened input tokens, it can be summarized at different explanation levels. First, to obtain a variable-specific temporal importance map, $\hat{\alpha}_{Time}$, $\hat{\alpha}_{TVA}$ is averaged over the forecasting horizon and reshaped into an $F \times T$ matrix, as in (16). Second, to obtain a compact variable importance vector, $\hat{\alpha}_{Var}$, $\hat{\alpha}_{Time}$ is averaged over the lag dimension, as in (17). Thus, $\hat{\alpha}_{TVA}$ provides the most detailed forecast-step-level explanation, $\hat{\alpha}_{Time}$ summarizes lag-wise importance for each variable, and $\hat{\alpha}_{Var}$ summarizes overall variable-level importance.

Fig. 3 visualizes these three explanation levels for a single predicted variable. Since DLFormer defines tokens at the variable-lag level, the TVA attention map can be interpreted as a model-internal saliency signal that reflects the relative predictive importance of each input region. However, attention weights should not be regarded as strict causal explanations. Therefore, in this study, the TVA attention map is interpreted not as evidence of causal effects between past variables and future predictions, but as an importance signal for variable-lag input regions used by DLFormer during forecasting.

IV. EXPERIMENTS

A. Dataset and Experimental Settings

We conducted comprehensive experiments to evaluate and analyze the predictive performance and explainability of DLFormer. For comparative analysis, we utilized one proprietary multivariate time-series dataset (Port Air Quality (PAQ)) and nine publicly available datasets (Volatility, PM2.5, SML, Exchange, Weather, ETTh1, ETTh2, ETTm1, and ETTm2). Table III provides detailed descriptions and variable information for each dataset.

All datasets were divided into training, validation, and test sets with a 70%, 10%, and 20% split, respectively. The mean and standard deviation computed from the training set were

TABLE III
DETAILS OF TIME-SERIES DATASETS USED IN THE EXPERIMENTS

Dataset	Number of Variables	Data Description	Number of Rows	Collected Period	Frequency
PAQ	7	Air Quality Datasets Collected from Busan Port (Non-Public)	9357	2004.03.01–2005.04.01	Hourly
SML	16	Public Datasets Collected from Home Monitoring Systems, used in [37], [46]	4137	2012.03.01–2014.05.01	15-minute
Volatility	8	Public dataset containing daily realized volatility values of the S&P 500 stock index, calculated from intraday data, along with its daily returns, used in [29].	4641	2000.01.01–2018.06.01	Daily
PM2.5	7	Public Meteorological Datasets Related to Fine Dust Collected in Beijing, China, used in [37]	33405	2010.01.01–2014.01.01	Hourly
Exchange	8	Public Exchange Rate Datasets for Specific Countries, used in the Darts Library [47]	7588	1990.01.01–2016.01.01	Daily
Weather	13	Public Outdoor Weather-Related Datasets Collected in Germany, used in the Darts Library [47]	52704	2020.01.01–2020.12.01	10-minute
ETTh1, 2	7	Public Datasets for Power Transformers in Substations, used in the Darts Library [47]	17420	2016.07.01–2018.06.01	Hourly
ETTm1, 2	7	Public Datasets for Power Transformers in Substations, used in the Darts Library [47]	69680	2016.07.01–2018.06.01	15-minute

used to normalize all splits, including the validation and test sets. In addition, no missing values were present in any of the datasets used in our experiments. To evaluate model performance, experiments were conducted on long-term forecasting tasks (96, 192, 336, and 720) and short-term forecasting tasks (3, 6, 12, and 36). The input sequence lengths were set to 48 and 36 for long-term and short-term forecasting, respectively.

Baseline We selected seven explainable forecasting models as baselines, covering both UDE and MDE approaches. The UDE baselines were TACN, NLinear, and iTransformer, whereas the MDE baselines included IMV-LSTM, XCM, MTEX, and TFT. Meanwhile, other post-hoc explanation methods, except for XCM and MTEX, were not included as direct comparison baselines. This is because their explanatory results can vary depending on the underlying forecasting model, and their explanation outputs and evaluation procedures are not directly comparable to those of existing MDE-based forecasting models. Therefore, our comparative evaluation was restricted to methods in which the forecasting architecture and the explanation mechanism are explicitly integrated or closely coupled.

Implementation Details The seven benchmark models were evaluated using their official GitHub implementations. Table IV presents the parameter setting ranges for each model.

The optimal parameters that yielded the best performance were selected and used in the comparative experiment. The L2 loss function was employed during training with a fixed batch size of 32. The Adam [43] optimizer was used for weight updates over 1000 training epochs. The model demonstrating the best performance on the validation dataset was applied to the test dataset, and the results were recorded. Additionally, to evaluate the average forecasting performance of each model, 10 independent training runs were conducted.

Although MTEX and XCM were originally developed for multivariate time-series classification, we adapted both models for forecasting to enable a fair comparison with MDE post-hoc explainability methods. Specifically, the classification projection layer was replaced with a forecasting projection layer whose output dimension corresponds to the forecasting horizon, and the Softmax activation function was removed. To ensure optimization fairness, the adapted MTEX and XCM models followed the same forecasting objective, data split, normalization protocol, and hyperparameter selection procedure as the other baselines. The checkpoint with the best validation performance was then used for test evaluation.

TABLE IV
HYPERPARAMETER SETTINGS FOR EACH MODEL

Models	Hyperparameter	Range
DLFormer	Number of Encoder, Decoder	{1, 2, 3}
	Size of embedding dimension	{128, 256, 512}
	Number of attention head	{2, 4, 8}
iTransformer	Number of Encoder	{1, 2, 3}
	Size of hidden dimension	{128, 256, 512}
NLinear	No Hyperparameter	-
TFT	Number of LSTM layer	{2, 3, 4}
	Size of hidden dimension	{64, 128, 256}
	Number of attention head	{4, 6, 8}
TACN	Number of TCN layer	{5, 6, 7}
	Size of TCN filter	{32, 64, 128}
IMV-LSTM	Size of hidden dimension	{64, 128, 256}
MTEX	Size of CNN filter	{64. 128, 256}
XCM	Size of CNN filter	{64. 128, 256}

B. Evaluation Metrics

We compared our model with each baseline model based on 1) forecasting performance and 2) explainability performance.

For predictive performance, we employed three evaluation metrics: mean squared error (MSE), mean absolute error (MAE), and correlation coefficient (COR). Their mathematical definitions are as follows:

$$MSE = \frac{1}{S}\sum_{s=1}(X - \hat{X})^2, \tag{18}$$

$$MAE = \frac{1}{S}\sum_{s=1}\left|X - \hat{X}\right|, \tag{19}$$

$$COR = \frac{\sum_{s=1}(X-\bar{X})(\hat{X}-\bar{\hat{X}})}{\sqrt{\sum_{s=1}(X-\bar{X})^2}\sqrt{\sum_{s=1}^{S}(\hat{X}-\bar{\hat{X}})^2}}. \tag{20}$$

Here, s denotes the index of time series sample S. Lower MSE and MAE values indicate more accurate predictions, while a higher COR value signifies better predictive performance. For explainability performance, we assessed the models from two perspectives: quantitative evaluation (robustness, faithfulness) and qualitative evaluation (validity).

The quantitative evaluation consists of two complementary aspects: robustness and faithfulness. Based on evaluation techniques designed in prior studies [44], [45], the protocol is organized as follows.

Robustness evaluates the stability of explanations generated by independently trained models under identical experimental conditions: 1) all trained forecasting models extract variable-level VI scores from the test set. 2) The variability of the extracted VI scores was measured using percentage-based metrics (standard deviation-STD and coefficient of variation-CV), and rank-based metrics (Kendall's tau-TAU and COR), to assess score variability and ranking consistency. Lower variability and higher ranking consistency of VI scores indicate greater robustness of the explanation process.

Faithfulness evaluates how faithfully the explanations provided by the TVA attention map reflect the actual predictive behavior of DLFormer: 1) the global TVA attention map was computed from the training set and averaged over the sample and prediction-horizon axes to construct a variable-by-lag explanation map. 2) For each variable, a sliding window was applied along the lag dimension, and top-ranked, bottom-ranked, and randomly selected windows were selected based on the average attention value within each window. 3) The selected windows were masked in the test input, and the resulting change in forecasting performance was measured. If the explanation faithfully identifies important inputs required for prediction, top-ranked window masking should result in larger performance degradation. For masking, the selected regions were replaced with the local mean computed within the corresponding test sample, which was intended to prevent out-of-distribution artifacts caused by zero masking.

For qualitative evaluation, the following procedure was applied. 1) The trained models extracted feature importance scores by considering both temporal and variable perspectives.

TABLE V
RESULTS OF COMPARATIVE EXPERIMENTS ON LONG-TERM TIME SERIES FORECASTING

Dataset		PAQ			SML			Volatility			PM2.5			Exchange		
Metrics		MSE	MAE	COR	MSE	MAE	COR	MSE	MAE	COR	MSE	MAE	COR	MSE	MAE	COR
Ours	96	**1.032***	**0.708***	**0.673***	0.275	0.406	0.903	0.303	0.412	0.309	**0.658***	**0.590***	**0.283***	**0.058**	**0.186**	**0.267**
	192	**1.072***	**0.735***	**0.633***	0.445	0.513	0.795	**0.549**	**0.574**	**0.360**	**0.702***	**0.614***	**0.172***	**0.129**	0.285	0.290
	336	**1.162**	**0.757***	**0.616***	**0.477**	**0.550**	0.755	1.552	0.900	0.374	**0.731**	**0.619***	**0.131***	0.266	0.401	0.276
	720	**1.136**	**0.759**	0.525	-	-	-	-	-	-	**0.738***	**0.628***	**0.095**	**0.256**	**0.399**	**0.688**
iTransformer	96	1.148	0.750	0.594	0.255	0.403	**0.920**	**0.285**	0.409	0.157	0.878	0.673	0.207	0.069	0.199	0.029
	192	1.191	0.783	0.555	0.381	0.509	**0.859**	0.644	0.630	0.125	0.936	0.701	0.144	0.135	**0.277**	0.000
	336	1.203	0.796	0.548	0.566	0.611	**0.798**	1.426	0.902	0.084	0.961	0.705	0.099	0.245	0.375	-0.030
	720	1.176	0.815	**0.546**	-	-	-	-	-	-	0.856	0.672	0.064	0.725	0.677	-0.020
NLinear	96	1.281	0.827	0.545	**0.249**	**0.399**	0.916	0.291	**0.406**	0.211	0.947	0.700	0.181	0.075	0.208	-0.070
	192	1.319	0.846	0.515	**0.372**	**0.497**	0.857	0.628	0.615	0.000	1.030	0.742	0.107	0.151	0.293	-0.110
	336	1.285	0.834	0.522	0.568	0.604	0.786	**1.399**	**0.890**	-0.020	1.028	0.742	0.072	0.282	0.402	-0.140
	720	1.289	0.857	0.497	-	-	-	-	-	-	0.884	0.697	0.056	0.788	0.692	-0.080
TFT	96	1.349	0.871	0.585	0.814	0.748	0.017	0.298	0.418	0.025	0.756	0.630	0.139	0.088	0.231	0.058
	192	1.405	0.915	0.486	0.991	0.817	0.000	0.612	0.613	0.234	0.784	0.641	0.078	0.173	0.322	-0.060
	336	1.344	0.883	0.550	1.131	0.862	-0.030	3.064	1.293	-0.470	0.797	0.644	0.051	**0.203**	**0.349**	0.226
	720	1.476	0.881	0.443	-	-	-	-	-	-	0.766	0.658	0.024	1.031	0.827	0.345
TACN	96	1.348	0.851	0.321	0.381	0.496	0.713	0.574	0.599	-0.230	0.703	0.640	0.186	0.084	0.230	0.100
	192	1.396	0.895	0.445	0.669	0.657	0.430	7.302	2.342	0.337	0.717	0.654	0.147	0.166	0.333	0.234
	336	1.301	0.838	0.447	0.839	0.742	0.374	5.918	1.972	0.151	0.737	0.662	0.107	0.313	0.459	0.107
	720	1.483	0.878	0.414	-	-	-	-	-	-	0.761	0.668	0.056	3.910	1.718	0.015
IMV-LSTM	96	1.355	0.928	0.520	0.616	0.630	0.903	18.11	4.056	**0.376**	0.672	0.689	0.246	0.759	0.761	0.260
	192	1.286	0.918	0.473	0.834	0.756	0.808	16.36	3.778	0.312	0.782	0.712	0.139	0.986	0.859	**0.374**
	336	1.202	0.831	0.427	1.178	0.890	0.754	24.07	4.604	**0.430**	0.741	0.661	0.113	1.267	0.950	**0.401**
	720	1.212	0.806	0.323	-	-	-	-	-	-	0.754	0.678	0.062	1.676	1.105	0.674
MTEX	96	1.308	0.870	0.467	0.597	0.624	0.897	13.20	3.401	0.090	0.725	0.671	0.009	0.579	0.539	-0.030
	192	1.449	0.950	0.460	1.416	0.950	0.603	10.00	2.544	0.163	0.721	0.666	0.036	2.858	1.332	-0.070
	336	1.412	0.940	0.457	1.712	1.037	0.712	20.56	4.256	0.054	0.739	0.673	0.027	1.893	1.040	-0.010
	720	1.876	0.974	0.065	-	-	-	-	-	-	0.758	0.670	0.002	3.301	1.678	0.327
XCM	96	1.312	0.867	0.362	0.705	0.657	0.803	6.873	1.841	0.150	0.691	0.628	0.230	0.315	0.418	0.091
	192	1.308	0.876	0.485	1.316	0.945	0.728	8.407	2.398	-0.160	0.724	0.652	0.140	1.276	0.981	0.093
	336	1.309	0.873	0.388	1.430	0.964	0.589	19.58	3.860	0.054	0.748	0.672	0.091	1.501	1.096	0.043
	720	1.748	0.975	0.166	-	-	-	-	-	-	0.758	0.676	0.026	3.832	1.699	0.589

* indicates statistical significance at $p < 0.05$ based on the Wilcoxon signed-rank test against the strongest baseline.

TABLE VI
RESULTS OF COMPARATIVE EXPERIMENTS ON LONG-TERM TIME SERIES FORECASTING

Dataset		Weather			ETTh1			ETTm1			ETTh2			ETTm2		
Metrics		MSE	MAE	COR	MSE	MAE	COR	MSE	MAE	COR	MSE	MAE	COR	MSE	MAE	COR
Ours	96	0.124	0.259	0.473	0.117	0.258	0.202	0.038	0.147	**0.284**	0.135	0.282	0.680	**0.064***	**0.184***	**0.725***
	192	0.209	0.353	**0.476**	0.161	0.315	0.156	0.093	0.228	0.210	**0.154***	**0.311***	0.597	**0.091***	**0.227***	**0.685**
	336	0.340	0.447	**0.450**	0.181	0.333	0.176	0.139	0.285	0.205	**0.182***	**0.345***	0.572	**0.123***	**0.271***	**0.641**
	720	0.408	0.470	0.320	0.137	0.296	**0.259**	0.216	0.377	0.172	**0.190***	**0.360***	0.515	**0.154***	**0.313***	0.588
iTransformer	96	0.121	0.247	0.469	0.058	0.185	0.261	**0.029**	**0.127**	0.281	**0.133**	**0.280**	**0.698**	**0.074**	0.197	0.712
	192	0.150	**0.281**	0.458	0.079	**0.215**	**0.224**	**0.044**	**0.160**	0.244	0.181	0.333	**0.647**	0.110	0.247	0.676
	336	**0.217**	**0.338**	0.413	**0.093**	**0.237**	**0.221**	**0.059**	**0.187**	**0.208**	0.231	0.381	**0.613**	0.141	0.287	0.634
	720	**0.318**	**0.415**	**0.439**	**0.090**	**0.237**	0.223	**0.078**	**0.215**	0.219	0.252	0.403	**0.558**	0.181	0.331	**0.623**
NLinear	96	0.137	0.263	0.373	**0.057**	**0.182**	0.262	0.031	0.131	0.235	0.134	0.281	0.658	0.083	0.211	0.675
	192	0.163	0.291	0.372	**0.078**	**0.215**	0.214	0.046	0.162	0.207	0.185	0.335	0.595	0.115	0.253	0.628
	336	0.231	0.347	0.341	0.096	0.243	0.196	0.060	0.188	0.172	0.241	0.389	0.557	0.146	0.290	0.580
	720	0.338	0.429	0.404	0.106	0.257	0.157	0.082	0.220	0.159	0.305	0.447	0.485	0.188	0.335	0.557
TFT	96	0.109	**0.238**	**0.498**	0.092	0.232	0.191	0.047	0.163	0.217	0.295	0.434	0.593	0.110	0.242	0.631
	192	0.160	0.304	0.475	0.101	0.246	0.102	0.089	0.232	0.059	0.310	0.455	0.529	0.266	0.399	0.579
	336	0.304	0.422	0.324	0.150	0.306	0.124	0.140	0.289	0.010	0.384	0.506	0.362	0.699	0.670	0.220
	720	0.422	0.500	0.154	0.344	0.484	-0.030	0.113	0.259	0.028	0.920	0.823	0.333	0.393	0.520	0.027
TACN	96	0.128	0.267	0.402	0.238	0.413	0.065	0.054	0.177	0.215	0.384	0.503	0.192	0.111	0.255	0.568
	192	0.169	0.318	0.458	0.297	0.468	0.081	0.094	0.233	0.140	1.173	0.853	0.131	0.213	0.360	0.432
	336	0.269	0.404	0.394	0.222	0.389	0.177	0.203	0.376	0.039	1.251	0.946	0.231	0.632	0.657	0.184
	720	0.371	0.481	0.363	0.361	0.513	-0.050	0.334	0.493	0.022	0.974	0.861	-0.020	1.960	1.156	0.158
IMV-LSTM	96	0.173	0.312	0.440	0.447	0.569	0.245	0.165	0.326	0.224	1.374	0.996	0.274	0.280	0.415	0.536
	192	0.240	0.368	0.453	0.768	0.787	0.154	0.315	0.474	0.188	1.785	1.170	0.194	0.262	0.395	0.444
	336	0.351	0.455	0.402	0.618	0.691	0.175	0.511	0.633	0.191	2.170	1.323	0.255	0.673	0.675	0.372
	720	0.444	0.510	0.318	1.190	1.013	0.244	1.194	1.027	0.154	1.707	1.157	0.285	2.051	1.256	0.194
MTEX	96	**0.107**	0.252	0.464	0.182	0.347	**0.331**	0.331	0.375	0.275	0.551	0.579	0.367	0.448	0.478	0.431
	192	**0.146**	0.293	0.468	0.165	0.325	0.108	0.098	0.234	**0.321**	0.495	0.559	0.472	0.350	0.440	0.517
	336	0.226	0.370	0.442	0.384	0.467	0.102	0.483	0.529	0.200	0.362	0.483	0.546	0.457	0.523	0.433
	720	0.476	0.538	0.316	0.214	0.376	0.029	0.420	0.500	0.165	0.448	0.529	0.335	0.628	0.627	0.263
XCM	96	0.109	0.246	0.464	0.146	0.307	0.197	0.055	0.177	0.281	0.331	0.459	0.570	0.144	0.287	0.704
	192	0.158	0.305	0.467	0.163	0.313	0.141	0.094	0.237	0.207	0.399	0.509	0.540	0.216	0.356	0.644
	336	0.243	0.376	0.394	0.145	0.335	0.154	0.140	0.295	0.144	0.375	0.496	0.477	0.304	0.431	0.584
	720	0.381	0.488	0.391	0.231	0.386	0.133	0.177	0.335	**0.284**	0.384	0.500	0.434	0.414	0.510	0.506

TABLE VII
RESULTS OF COMPARATIVE EXPERIMENTS ON SHORT-TERM TIME SERIES FORECASTING

Dataset		PAQ			SML			Volatility			PM2.5			Exchange		
Metrics		MSE	MAE	COR	MSE	MAE	COR	MSE	MAE	COR	MSE	MAE	COR	MSE	MAE	COR
Ours	3	0.432	**0.435**	0.543	**0.000**	0.015	0.944	0.022	0.103	0.018	**0.101**	**0.189***	**0.205***	**0.003**	**0.037**	**0.066**
	6	0.581	**0.511**	0.667	**0.001**	**0.025**	0.952	0.034	0.129	0.059	**0.181**	**0.266***	**0.245***	**0.005**	**0.051**	0.088
	12	**0.684**	**0.562***	0.691	0.007	0.050	0.953	0.070	0.193	0.030	**0.310**	**0.356***	**0.301***	**0.009**	**0.069**	0.102
	36	**0.880***	**0.643***	**0.702***	0.090	0.196	0.919	0.140	0.277	**0.145**	**0.534***	**0.509**	**0.291**	0.026	0.121	**0.189**
iTransformer	3	**0.423**	0.449	**0.556**	0.001	0.017	**0.950**	0.020	0.098	0.041	0.118	0.207	0.131	**0.003**	0.040	0.020
	6	**0.564**	0.525	**0.669**	0.002	0.026	**0.955**	0.035	0.129	0.040	0.209	0.283	0.164	**0.005**	0.052	0.029
	12	0.686	0.582	**0.700**	**0.006**	**0.048**	**0.956**	0.060	0.172	0.064	0.352	0.382	0.214	**0.009**	0.070	0.027
	36	0.919	0.669	0.663	**0.067**	**0.173**	**0.946**	0.133	0.272	0.089	0.662	0.557	0.213	**0.024**	**0.119**	0.010
NLinear	3	0.516	0.494	0.502	**0.000**	**0.013**	0.922	**0.019**	**0.095**	-0.010	0.122	0.206	0.051	**0.003**	0.042	0.034
	6	0.683	0.575	0.634	**0.001**	0.026	0.906	**0.032**	0.127	-0.020	0.229	0.290	0.095	**0.005**	0.054	-0.010
	12	0.819	0.628	0.653	0.010	0.064	0.883	**0.054**	**0.167**	-0.010	0.400	0.400	0.147	0.010	0.072	-0.020
	36	1.053	0.718	0.611	0.156	0.270	0.844	**0.128**	**0.265**	0.074	0.727	0.580	0.157	0.026	0.123	-0.040
TFT	3	0.471	0.467	0.503	0.001	0.019	0.898	0.026	0.113	-0.020	0.113	0.203	0.140	**0.003**	0.040	-0.040
	6	0.718	0.592	0.668	0.009	0.060	0.897	0.046	0.149	0.032	0.216	0.287	0.183	**0.005**	0.055	-0.030
	12	0.802	0.621	0.677	0.016	0.090	0.899	0.059	0.178	0.000	0.360	0.393	0.231	**0.009**	0.071	0.029
	36	1.415	0.863	0.661	0.305	0.386	0.768	0.176	0.321	0.007	0.609	0.551	0.186	0.031	0.138	-0.100
TACN	3	0.514	0.491	0.364	0.002	0.035	0.418	0.020	0.097	-0.020	0.123	0.215	0.069	0.004	0.043	-0.010
	6	0.696	0.586	0.437	0.008	0.065	0.391	0.033	**0.123**	-0.040	0.220	0.299	0.141	0.006	0.057	-0.030
	12	0.839	0.651	0.463	0.025	0.104	0.712	0.058	0.168	0.021	0.352	0.398	0.231	0.021	0.090	0.000
	36	1.098	0.747	0.522	0.177	0.294	0.692	0.132	0.269	0.104	0.572	0.547	0.202	0.027	0.127	-0.010
IMV-LSTM	3	0.557	0.526	0.529	0.030	0.118	0.914	8.606	2.626	**0.101**	0.105	0.199	0.193	0.245	0.383	0.060
	6	0.790	0.623	0.610	0.057	0.158	0.909	11.51	3.130	**0.135**	0.192	0.284	0.236	0.287	0.418	**0.101**
	12	0.978	0.705	0.608	0.072	0.194	0.903	13.01	3.353	**0.099**	0.313	0.378	0.286	0.390	0.519	**0.126**
	36	1.311	0.840	0.579	0.231	0.360	0.916	17.01	3.903	0.127	0.551	0.515	0.279	0.723	0.733	0.187
MTEX	3	0.724	0.563	0.370	0.006	0.060	0.890	10.95	2.647	-0.060	0.286	0.338	0.113	0.240	0.298	0.031
	6	0.909	0.654	0.512	0.014	0.082	0.869	11.71	2.911	-0.060	0.265	0.336	0.187	0.305	0.357	0.072
	12	1.034	0.713	0.507	0.423	0.346	0.710	8.556	2.284	-0.030	0.368	0.416	0.215	0.301	0.342	0.081
	36	1.270	0.823	0.444	0.162	0.305	0.869	9.513	2.340	-0.210	0.606	0.585	0.177	0.646	0.562	0.007
XCM	3	0.687	0.603	0.403	0.019	0.105	0.508	0.490	0.592	0.070	0.119	0.227	0.145	0.237	0.367	0.006
	6	0.909	0.711	0.465	0.050	0.166	0.662	0.390	0.669	0.050	0.203	0.301	0.185	0.236	0.312	0.010
	12	1.045	0.750	0.492	0.051	0.163	0.811	0.382	0.489	0.009	0.341	0.410	0.236	0.155	0.264	0.022
	36	1.232	0.830	0.531	0.187	0.326	0.866	0.690	0.653	-0.190	0.561	0.556	0.270	0.097	0.239	0.046

* indicates statistical significance at $p < 0.05$ based on the Wilcoxon signed-rank test against the strongest baseline.

TABLE VIII
RESULTS OF COMPARATIVE EXPERIMENTS ON SHORT-TERM TIME SERIES FORECASTING

Dataset		Weather			ETTh1			ETTm1			ETTh2			ETTm2		
Metrics		MSE	MAE	COR	MSE	MAE	COR	MSE	MAE	COR	MSE	MAE	COR	MSE	MAE	COR
Ours	3	**0.001**	**0.023**	**0.306**	**0.007**	**0.061**	**0.208**	**0.002**	**0.032**	**0.096**	**0.007***	**0.057***	**0.771***	**0.001**	**0.016**	**0.801***
	6	**0.003**	**0.036**	**0.353**	**0.011**	0.080	0.251	**0.003**	**0.041**	**0.114**	**0.020***	**0.094***	**0.782***	**0.002**	**0.026***	0.724
	12	0.009	**0.062**	0.355	0.021	0.110	0.303	**0.006**	**0.055**	**0.151**	**0.041***	**0.144***	**0.822***	**0.008**	**0.055**	**0.711**
	36	**0.030***	**0.118***	0.338	0.046	0.165	0.306	0.025	0.118	**0.258**	**0.080***	**0.214***	**0.751**	**0.047***	**0.138***	**0.743***
iTransformer	3	**0.001**	0.024	0.304	**0.007**	**0.061**	0.201	**0.002**	0.033	0.054	0.011	0.070	0.740	**0.001**	**0.019**	0.771
	6	**0.003**	0.037	0.341	0.012	**0.079**	**0.271**	0.004	0.042	0.092	0.029	0.113	0.744	**0.002**	**0.031**	**0.727**
	12	0.009	**0.062**	**0.356**	**0.019**	**0.104**	**0.327**	**0.006**	0.056	0.130	0.055	0.164	0.796	**0.008**	0.056	0.707
	36	0.057	0.152	**0.339**	0.037	0.147	0.292	**0.016**	**0.093**	0.206	0.089	0.223	0.744	0.053	0.144	0.707
NLinear	3	**0.001**	0.024	0.258	**0.007**	0.062	0.150	**0.002**	**0.032**	0.019	0.013	0.074	0.693	0.002	0.026	0.636
	6	**0.003**	0.037	0.300	0.012	0.081	0.224	0.004	0.042	0.028	0.033	0.118	0.697	0.004	0.041	0.567
	12	0.010	0.063	0.304	0.020	**0.104**	0.317	0.007	0.057	0.054	0.058	0.169	0.755	0.014	0.075	0.403
	36	0.061	0.155	0.214	**0.036**	**0.144**	0.314	0.017	0.095	0.163	0.090	0.223	0.719	0.096	0.204	0.488
TFT	3	**0.001**	**0.023**	**0.306**	0.009	0.070	0.167	**0.002**	0.034	0.055	0.016	0.074	0.720	**0.001**	**0.019**	0.728
	6	0.004	0.048	0.218	0.016	0.095	0.228	0.004	0.045	0.070	0.035	0.116	0.766	0.003	0.032	0.663
	12	0.012	0.069	0.323	0.030	0.130	0.296	0.008	0.060	0.142	0.072	0.188	0.783	0.012	0.066	0.589
	36	0.060	0.158	0.143	0.083	0.223	**0.335**	0.025	0.118	0.145	0.193	0.344	0.712	0.082	0.192	0.538
TACN	3	0.002	0.029	0.165	0.010	0.078	0.168	0.003	0.037	0.008	0.018	0.098	0.535	0.002	0.030	0.388
	6	0.005	0.045	0.210	0.020	0.102	0.190	0.005	0.053	0.050	0.035	0.131	0.584	0.004	0.043	0.350
	12	0.009	0.065	0.193	0.031	0.138	0.155	0.009	0.068	0.093	0.065	0.187	0.700	0.030	0.111	0.316
	36	0.056	0.166	0.228	0.134	0.305	0.043	0.026	0.119	0.208	0.134	0.285	0.593	0.110	0.228	0.421
IMV-LSTM	3	0.003	0.027	0.305	0.015	0.092	0.138	0.005	0.052	0.057	0.017	0.091	0.674	0.003	0.027	0.730
	6	0.005	0.045	0.321	0.030	0.132	0.151	0.013	0.085	0.057	0.062	0.177	0.604	0.005	0.049	0.617
	12	0.012	0.076	0.314	0.077	0.215	0.258	0.023	0.118	0.092	0.104	0.238	0.631	0.016	0.086	0.503
	36	0.092	0.214	0.292	0.186	0.342	0.226	0.087	0.227	0.219	0.328	0.446	0.529	0.144	0.273	0.491
MTEX	3	0.004	0.036	0.004	0.013	0.080	0.011	0.005	0.049	0.018	0.019	0.099	0.530	0.004	0.048	0.145
	6	0.005	0.049	0.021	0.021	0.104	0.165	0.007	0.057	0.012	0.190	0.251	0.593	0.008	0.064	0.294
	12	0.015	0.095	0.252	0.036	0.142	0.300	0.012	0.076	0.016	0.092	0.218	0.754	0.199	0.228	0.316
	36	0.042	0.146	0.308	0.290	0.357	0.252	0.026	0.119	0.221	0.306	0.407	0.579	0.414	0.431	0.310
XCM	3	0.003	0.027	0.245	0.023	0.116	0.134	0.005	0.052	0.090	0.025	0.117	0.593	0.003	0.035	0.598
	6	0.005	0.046	0.315	0.031	0.133	0.204	0.008	0.065	0.101	0.062	0.186	0.656	0.006	0.056	0.504
	12	**0.008**	0.064	0.320	0.044	0.160	0.250	0.014	0.085	0.149	0.109	0.246	0.692	0.018	0.094	0.556
	36	0.047	0.156	0.305	0.076	0.213	0.273	0.031	0.129	0.236	0.203	0.352	0.641	0.089	0.216	0.619

* indicates statistical significance at $p < 0.05$ based on the Wilcoxon signed-rank test against the strongest baseline.

TABLE IX
OVERALL AVERAGE RANKING OF FORECASTING RESULTS

Models	Long	Short	**AVG**
Ours	**2.475**	1.650	**2.062**
iTransformer	2.388	**2.337**	2.362
NLinear	3.163	3.125	3.144
TFT	4.425	4.213	4.319
TACN	5.338	4.750	5.044
IMV-LSTM	6.850	6.513	6.681
MTEX	6.150	6.862	6.506
XCM	5.213	6.550	5.881

2) These scores were compared with domain knowledge and prior studies associated with each dataset to evaluate the validity and interpretability of the model explanations.

C. Experimental Results

Forecasting Performance Tables V and VI present the overall results of the long-term forecasting experiments, while Tables VII and VIII show the results for the short-term forecasting tasks. All values are averaged over 10 independent runs. The best performance for each dataset and forecasting length is highlighted in bold red, whereas the second-best performance is underlined in blue. An asterisk (*) indicates that the proposed method achieves a statistically significant improvement over the strongest baseline according to the Wilcoxon signed-rank test ($p < 0.05$). Due to dataset size limitations, forecasting experiments of length 720 were not conducted for the SML and Volatility datasets.

MDE models exhibit lower forecasting performance compared to UDE models. In contrast, DLFormer consistently outperforms or matches UDE methods. In Table V, DLFormer achieves the best results most frequently, with the previously recognized state-of-the-art model, iTransformer, ranking second, followed by NLinear. In Table VI, DLFormer shows relatively weaker performance on the Weather, ETTh1, and ETTm1 datasets but delivers the best performance on the ETTh2 and ETTm2 datasets.

In long-term forecasting, DLFormer demonstrates significant improvements in average MSE over other models: 4.03 over iTransformer, 10.58 over NLinear, 49.12 over TFT, 152.14 over TACN, 500.25 over IMV-LSTM, 377.58 over MTEX, and 287.15% over XCM.

Similarly, in short-term forecasting (Table VII), DLFormer and iTransformer show comparable performance, consistently outperforming the other models. DLFormer achieves the highest number of best results, followed closely by NLinear. In Table VIII, DLFormer displays relatively weaker performance on the ETTh1 and ETTm1 datasets but outperforms other models on the remaining datasets, showing comparable performance to iTransformer. DLFormer improves average MSE by 0.79 over iTransformer, 17.09 over NLinear, 29.06 over TFT, 20.34 over TACN, 1143.59 over IMV-LSTM, 968.38 over MTEX, and 89.74 over XCM in short-term forecasting tasks.

TABLE X
ROBUSTNESS EVALUATION RESULTS FOR EXPLANATIONS: PERCENTAGE-BASED

Models	Ours		TFT		IMV-LSTM		MTEX		XCM	
Metrics	STD	CV	STD	CV	STD	CV	STD	CV	STD	CV
PAQ	1.624	0.111	7.351	0.501	3.473	0.242	**0.215**	**0.015**	1.855	0.137
SML	0.685	0.110	3.648	0.564	2.769	0.403	**0.027**	**0.004**	2.328	0.446
Volatility	3.085	0.240	7.164	0.562	3.196	0.252	**0.320**	**0.025**	4.615	0.396
PM2.5	2.047	0.149	5.480	0.445	2.645	0.194	**0.046**	**0.003**	4.320	0.317
Exchange	3.765	0.345	6.243	0.512	4.377	0.361	**0.174**	**0.013**	2.541	0.202
Weather	1.739	0.231	3.071	0.462	1.509	0.194	**0.016**	**0.002**	1.150	0.160
ETTh1	2.454	0.164	8.089	0.590	5.648	0.410	**0.062**	**0.004**	2.575	0.179
ETTm1	2.719	0.187	5.134	0.431	5.254	0.371	**0.033**	**0.002**	3.017	0.208
ETTh2	1.891	0.142	7.907	0.591	5.303	0.403	**0.537**	**0.037**	2.686	0.203
ETTm2	2.155	0.179	6.129	0.506	6.000	0.508	**0.112**	**0.078**	2.175	0.186

TABLE XI
ROBUSTNESS EVALUATION RESULTS FOR EXPLANATIONS: RANK-BASED

Models	Ours		TFT		IMV-LSTM		MTEX		XCM	
Metrics	TAU	COR	TAU	COR	TAU	COR	TAU	COR	TAU	COR
PAQ	**0.765**	**0.885**	0.121	0.158	0.206	0.252	0.299	0.368	0.348	0.428
SML	**0.720**	**0.857**	0.118	0.168	0.269	0.369	0.269	0.351	0.322	0.372
Volatility	0.374	**0.480**	0.195	0.251	0.126	0.165	0.157	0.200	**0.428**	0.429
PM2.5	**0.720**	**0.821**	0.316	0.395	0.303	0.404	0.482	0.561	0.301	0.340
Exchange	**0.625**	**0.777**	0.215	0.278	0.041	0.075	0.238	0.234	0.169	0.186
Weather	0.479	0.616	0.436	0.567	0.046	0.055	0.118	0.134	**0.557**	**0.635**
ETTh1	**0.722**	**0.813**	0.113	0.164	0.180	0.252	0.610	0.666	0.458	0.573
ETTm1	0.519	0.668	0.377	0.492	0.297	0.373	**0.731**	**0.840**	0.238	0.326
ETTh2	**0.657**	**0.769**	0.066	0.103	0.266	0.356	0.253	0.285	0.291	0.319
ETTm2	**0.598**	**0.707**	0.299	0.363	0.243	0.330	0.582	0.652	0.208	0.230

TABLE XII
ROBUSTNESS EVALUATION RESULTS FOR EXPLANATIONS: RANK-BASED

Models	Percentage	Rank	**AVG**
Ours	2.30	**1.25**	**1.77**
TFT	4.90	3.75	4.32
IMV-LSTM	3.75	4.00	3.87
MTEX	**1.00**	2.80	1.90
XCM	3.05	3.10	3.07

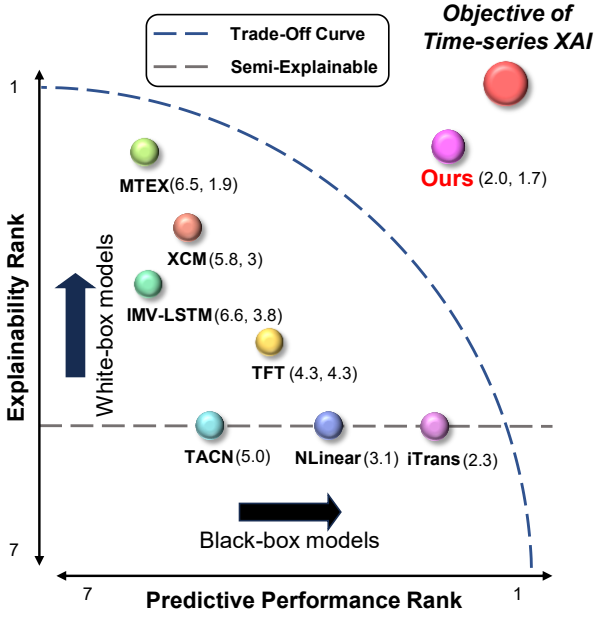

Fig. 4. Performance-explainability trade-off for time-series forecasting models.

Table IX presents the weighted average results for each dataset, ranked by long-term and short-term forecasting horizons, to quantitatively identify the best-performing models. The results confirm that DLFormer achieves the highest performance in long-term forecasting. The iTransformer ranks second, followed by NLinear, XCM, TFT, TACN, MTEX, and IMV-LSTM, in that order. In short-term forecasting, iTransformer ranks highest, closely followed by DLFormer, consistently outperforming the other models. Subsequent ranks include NLinear, TFT, TACN, IMV-LSTM, XCM, and MTEX, with
the remaining models showing comparable performance overall.

DLFormer shows consistently strong and stable performance in both long- and short-term forecasting, achieving top overall rankings. Despite being MDE-based, it delivers state-of-the-art results across various datasets. This highlights the effectiveness of the TVAL strategy, which prevents dimensional mixing in Transformer models and improves forecasting accuracy by capturing complex time and variable-wise dependencies.

Robustness Evaluation of Explainability Performance Tables X and XI present the robustness evaluation results for the explanations generated by each explainable time-series model across different datasets. Each value reflects the outcomes from applying 10 independently trained models to the complete test set. The robustness of VI scores, extracted from multi-dimensional explanations, is assessed using percentage- and rank-based metrics. For DLFormer, we used

TABLE XIII
RELATIVE MAE CHANGE (%) OF DLFORMER UNDER PERTURBATION-BASED VARIABLE-LAG MASKING.

	Masking ratio	5%			10%			20%		
	Masking strategy	Top	Random	Bottom	Top	Random	Bottom	Top	Random	Bottom
Long-term forecasting	PAQ	**-0.069**	0.002	0.224	**-2.407**	-0.281	-0.392	**-14.201**	0.054	-0.816
	SML	**-0.250**	-0.080	-0.037	**-0.642**	-0.241	-0.175	**-0.905**	0.187	-0.744
	Volatility	1.396	0.606	**0.310**	2.196	1.826	**0.195**	2.169	2.302	**-4.319**
	PM2.5	**0.033**	0.108	0.04	**-2.929**	0.164	0.042	**-3.268**	0.174	0.059
	Exchange	**-0.423**	0.046	0.212	**-1.273**	-0.739	-0.204	**-1.696**	0.192	0.092
	Weather	0.117	**0.001**	0.137	0.343	**-0.304**	0.037	**-1.775**	0.193	-0.112
	ETTh1	**-3.284**	-0.446	0.178	**-4.405**	-0.709	-0.157	**-4.480**	-1.707	-0.432
	ETTm1	**-0.177**	-0.147	1.088	**-0.163**	0.043	1.431	**-0.577**	0.08	1.885
	ETTh2	**-5.971**	-0.718	0.089	**-7.598**	-0.292	0.064	**-12.689**	-2.112	-0.030
	ETTm2	**-1.358**	-0.014	0.175	**-1.803**	-0.177	0.098	**-7.936**	-6.197	0.195
Short-term forecasting	PAQ	**-30.580**	-1.178	-0.336	**-49.443**	-2.184	-0.727	**-55.775**	-11.825	-1.124
	SML	**-946.914**	-2.420	0.525	**-968.005**	-0.653	0.335	**-1123.375**	-4.139	-1.076
	Volatility	0.145	**0.124**	0.452	**0.752**	0.882	0.787	-15.447	**-16.733**	1.190
	PM2.5	**-3.227**	-1.208	0.017	**-9.707**	-0.583	0.004	**-12.928**	-1.840	-0.145
	Exchange	**-0.013**	-0.005	0.016	-0.054	0.010	**-17.569**	-0.104	-0.003	**-13.164**
	Weather	**-59.296**	0.133	-0.008	**-67.589**	-0.619	-0.138	**-136.046**	-38.348	-0.386
	ETTh1	-0.049	0.042	**-0.072**	-0.103	-0.081	**-0.114**	**-26.579**	-0.050	0.158
	ETTm1	**-0.089**	0.037	0.202	**-0.193**	0.104	0.484	-0.392	0.022	**-11.498**
	ETTh2	**-183.437**	-2.120	-0.075	**-181.228**	-2.091	-0.177	**-174.597**	-7.562	-0.094
	ETTm2	**-53.240**	-14.735	0.074	**-72.946**	-0.155	0.109	**-75.398**	0.007	0.034

the $\hat{\alpha}_{var}$ described in Section III-C. The best results are highlighted in bold red and second-best results are underlined in blue. In Table X, MTEX demonstrates superior robustness in percentage-based VI, achieving the best results across all evaluation metrics. DLFormer mostly ranks second, exhibiting consistently high robustness.

In contrast, Table XI, which evaluates the consistency of variable rankings, shows that DLFormer achieves the best robustness in most cases. XCM ranks second, followed by MTEX, TFT, and IMV-LSTM in that order. From a quantitative perspective, the average variability of VI percentages across individually trained DLFormer models is 2.123%, indicating high stability. In terms of ranking consistency, the average COR between the individually trained DLFormer models reaches 67.6%, demonstrating remarkable robustness compared to most other models. Table XII summarizes the overall ranking for the quantitative evaluation of explainability, averaging rankings of each metric across all datasets. The results confirm that DLFormer achieves the best comprehensive performance in terms of explanation robustness. Thus, based on forecasting performance comparisons across various settings and quantitative explainability, DLFormer clearly outperforms existing explainable time-series methods in both aspects.

Fig. 4 illustrates the trade-off between forecasting performance and explainability. It compares where models stand in terms of both, with lower ranks being better. MDE-based models like MTEX show moderate forecasting but strong explainability. In contrast, UDE-based models like TACN are only evaluated on forecasting due to limited explainability. DLFormer stands out by achieving both high accuracy and strong explainability, effectively reducing the usual trade-off.

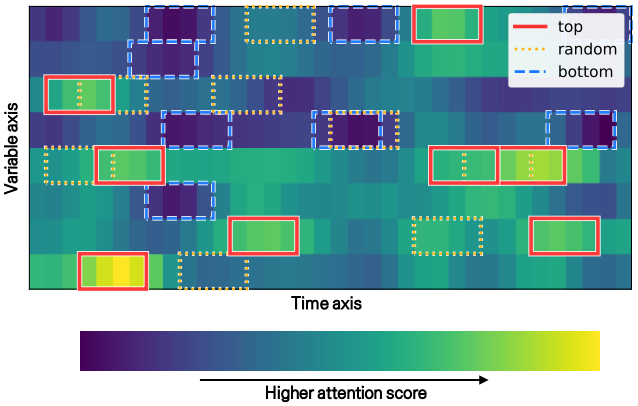

Fig. 5. Example of the perturbation-based variable-lag masking protocol. Top-ranked masking: Red solid boxes, random masking: yellow dotted boxes, bottom-ranked masking: blue dashed boxes.

Faithfulness Evaluation of Explainability To examine whether the TVA attention maps of the proposed DLFormer reflect input regions closely related to forecasting, we conducted a perturbation-based faithfulness evaluation using a variable-lag masking protocol. This protocol follows the procedure described in Section IV-B, and Fig. 5 provides an intuitive example of the selected variable-lag windows. Specifically, top, bottom, and random denote windows selected from high-attention, low-attention, and randomly sampled variable-lag regions, respectively. The main assumption is that, if the attention map faithfully highlights input regions that are important for prediction, masking high-attention variable-lag regions should degrade forecasting performance more than masking randomly selected or low-attention regions.

Table XIII reports the relative MAE change (%) compared with the non-masked DLFormer for both long-term and short-term forecasting. More negative values indicate larger

performance degradation after masking. The masking ratio is defined as the proportion of data points included in the selected windows relative to the total number of variable-lag data points. We evaluated masking ratios of 5%, 10%, and 20%.

Overall, masking the top-ranked variable-lag windows resulted in larger performance degradation than masking random or bottom-ranked windows in most datasets. Quantitatively, top-ranked masking produced, on average, 70.075 and 71.278 percentage points larger performance degradation than random and bottom masking, respectively. This indicates that high-attention regions in the TVA attention maps are generally aligned with input regions to which DLFormer's predictions are more sensitive.

This effect was more pronounced in short-term forecasting. In the short-term setting, the magnitude of degradation caused by top-window masking was substantially larger than that observed in the long-term setting. Specifically, top-window masking produced a relative MAE change of -2.468% in the long-term setting, whereas it caused a much larger change of -141.529% in the short-term setting. Even when using the median to reduce the influence of large relative values, the degradation remained stronger in short-term forecasting (-28.580%) than in long-term forecasting (-1.316%). This is reasonable because short-term predictions are more directly linked to local temporal evidence within the input window. In contrast, long-term forecasting tends to rely more heavily on distributed temporal patterns, accumulated uncertainty, and broader trend or seasonal information; therefore, removing a small number of local variable-lag windows can have a weaker or less monotonic effect. Nevertheless, top-window masking still produced the largest degradation in 48 out of 60 dataset-task-ratio cases.

The dataset-level results also reveal meaningful differences. Environmental and sensor-based datasets such as PAQ, PM2.5, SML, and Weather showed more pronounced degradation under top masking, suggesting that DLFormer relies on localized variable-lag evidence in these domains. The ETTm2 and ETTh2 datasets also exhibited clear perturbation sensitivity, indicating that the attention maps capture predictive temporal structures. In contrast, the weaker or less monotonic patterns observed in ETTh1, ETTm1, Exchange, and Volatility may reflect the more stochastic nature of these datasets, higher randomness, or weaker localization of the learned predictive evidence. This interpretation is also consistent with the forecasting performance results in Tables VI and VIII, where ETTh1 and ETTm1 show relatively weaker predictive performance than ETTh2 and ETTm2. In addition, Exchange and Volatility are generally known to contain relatively strong stochastic components. Therefore, these results should be interpreted not as strict causal identification, but as perturbation-based evidence of predictive faithfulness. Overall, the results suggest that the TVA attention maps faithfully reflect the general predictive behavior of DLFormer.

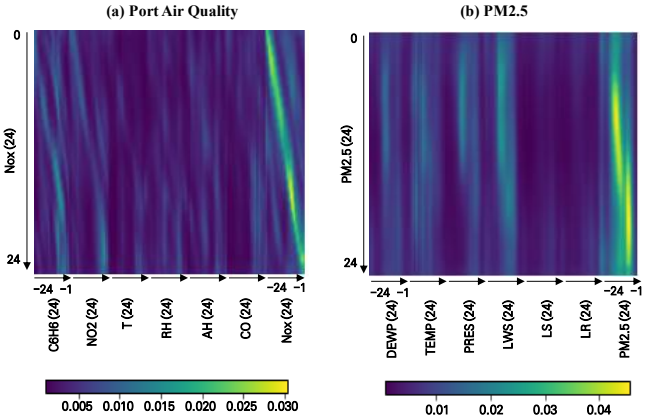

Fig. 6. Visualization of the time-variable-aware attention map of DLFormer for (a) Port Air Quality and (b) PM2.5 datasets.

Qualitative Evaluation of Explainability For the qualitative explainability evaluation, we utilized two datasets related to environmental pollution, PAQ and PM2.5, which are well-studied in the literature. These datasets facilitate qualitative comparisons based on domain knowledge. Fig. 6 presents the visualization of the DLFormer TVA attention maps for the two datasets. Fig. 6 visualizes feature importance extracted from these two datasets using five different models, analyzing from TI and VI perspectives. A detailed explanation of these results follows.

a) Port Air Quality Dataset Analysis: Air quality in port environments typically exhibits a 24-hour periodicity due to ongoing port operations [48], [49]. Ship arrivals, departures, and vehicular activities predominantly occur during the diurnal period, generating concentrated air pollution. Emissions from these activities interact and have a prolonged impact on air quality over subsequent hours [50]. As shown in Fig. 6, the TVA attention map for the PAQ dataset reveals that DLFormer consistently emphasizes the NOx predictor values from 24 hours prior at each prediction step.

This observation aligns well with domain knowledge regarding the daily periodicity of port emissions. Such periodicity is observed across all variables, indicating strong correlations driven by the cyclical nature of air quality in port operations.

In Fig. 7, the DLFormer results for the PAQ dataset, aggregated over forecasting horizons, reflect the general influence of past time points. While some noise exists, the model emphasizes time points from 10 to 13 hours prior, suggesting that daytime emissions persistently impact air quality later in the day. For VI, NOx exhibits the highest contribution due to autoregressive characteristics. NO2, a component of NOx, shows the second-highest importance. Additionally, C6H6 and CO demonstrate higher contributions than temperature and humidity, consistent with prior studies indicating strong correlations between these pollutants and NOx [51], [52].

VI results from IMV-LSTM appear partially reasonable, with NOx and CO receiving high importance. Additionally,

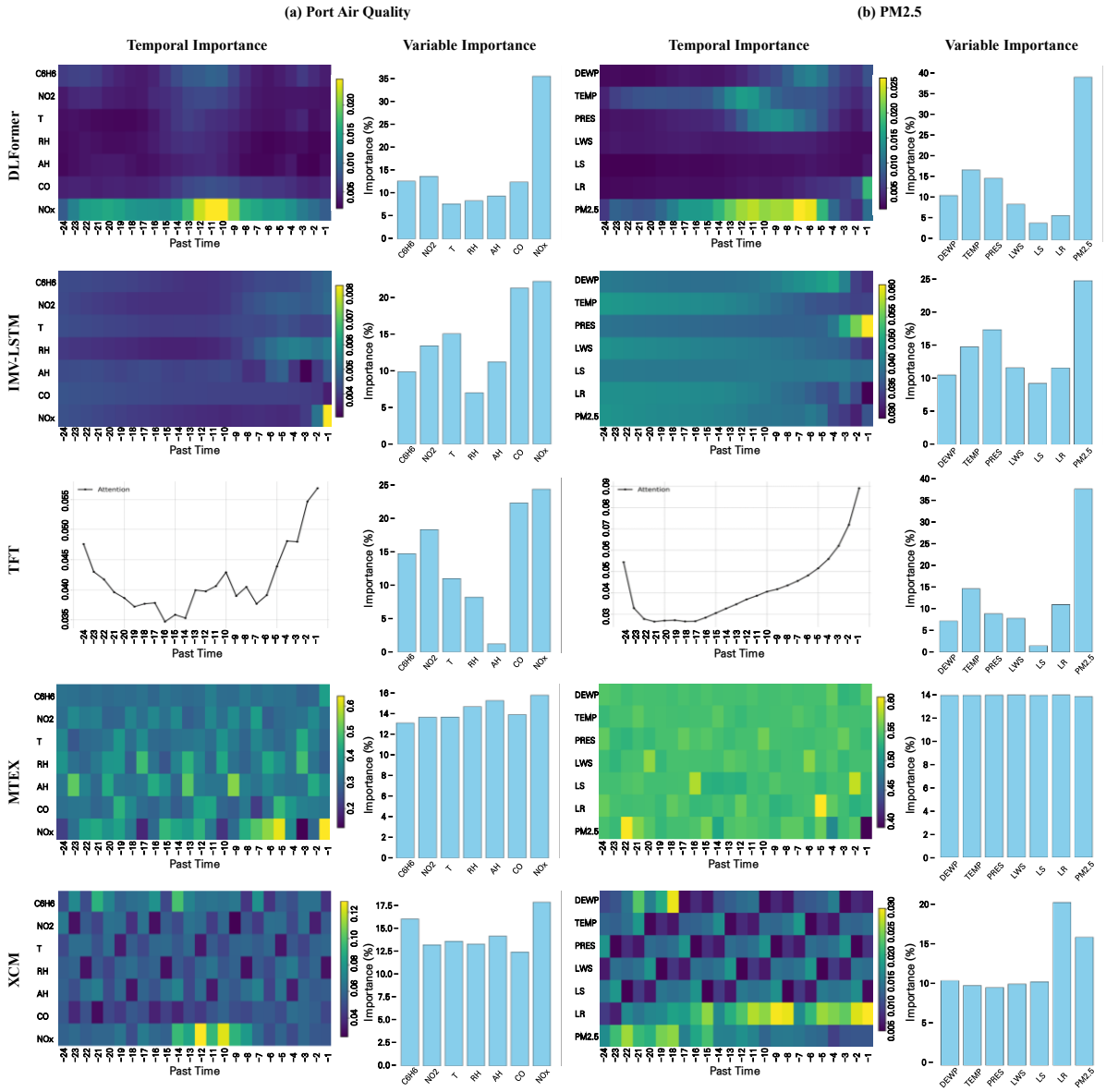

FIGURE 7. Visualization comparing temporal importance and variable importance across different models for two datasets: (a) Port Air Quality and (b) PM2.5.

while the model captures the stepwise importance of earlier time points, its RNN-based architecture has limitations in modeling long-range dependencies. As the time lag increases, RNNs tend to assign uniform or diminished importance to distant time steps, complicating the accurate reflection of temporal influences from the past.

TFT provides reasonable contributions from NOx, NO2, and CO. However, the relatively lower contribution of NO2 compared to CO is inconsistent with domain knowledge. TI effectively captures influences from 10 to 13 h prior, recent time points, and the 24-hour periodicity, demonstrating the model's strong explainability.

MTEX yields expected results, with NOx dominating the VI and recent time points dominating TI. However, the model fails to capture the 24-hour periodic pattern and shows occasional spikes in TI that are difficult to interpret from a domain perspective.

Finally, XCM captures the impact of pollutant emissions 10 to 13 hours prior but neglects periodicity and shows a lower contribution of NO2 compared to other features.

These results indirectly emphasize the superior forecasting performance of DLFormer on the PAQ dataset, as it effectively incorporates NO2, known for its strong correlation with NOx, into its predictions.

b) PM2.5 Dataset Analysis: In Beijing, PM2.5 concentrations and air quality indices are heavily influenced by meteorological conditions and human activities affecting atmospheric mixing. Thus, they typically exhibit a 24-hour

periodicity, with additional immediate effects from wind, rain, and snow [53]-[55].

In Fig. 7, the VI results from the DLFormer show that snow (LS) and rain (LR) generally have lower absolute contributions due to their limited occurrences in the dataset. However, recent time points exert relatively greater influence when these events occur. Temperature (TEMP), dew point temperature (DEWP), and pressure (PRES) significantly impact atmospheric stability and circulation, with their influence primarily concentrated within the recent 12-hour window, consistent with existing literature. However, cumulative wind speed (LWS) tends to be underestimated. In Fig. 5, the TVA attention map for the PM2.5 dataset further supports the presence of this 24-hour cycle, with the model focusing on the periodic component and recent time intervals. PM2.5 values from 24 hours prior receive heavy weighting, while other variables also emphasize recent hours, reflecting a balance between long-term periodicity and short-term effects. The VI results for IMV-LSTM indicate strong autoregressive contributions from PM2.5, similar to DLFormer, but with PRES ranking higher than TEMP, which contradicts domain knowledge. For TI, the model primarily captures short-term dependencies while largely neglecting the impact of LR.

TFT shows comparable VI patterns to DLFormer, with appropriate contributions from PM2.5, TEMP, and LR, while giving slightly higher importance to LR than DLFormer. In terms of TI, TFT effectively captures the 24-hour periodicity and recent contributions from LWS, LS, LR, and the target variable, demonstrating strong explainability.

For MTEX, VI results show uniform contributions across all variables, indicating limited differentiation. TI focuses on PM2.5 values from 22 hours prior, revealing partial periodicity. LS and LR show some recent influence, while other variables' temporal patterns are poorly distinguished.

Similarly, XCM captures the periodicity from approximately 22 hours prior and reflects a strong influence from recent LR values. However, it shows uniform temporal patterns across other variables, with LS, TEMP, and PRES exhibiting similar patterns contrary to domain expectations. Qualitative analysis demonstrates that DLFormer and TFT produce VI and TI patterns that are most consistent with domain knowledge across the two datasets. In contrast, IMV-LSTM exhibits partially consistent TI for short-term dependencies, but tends to converge toward the mean as the time distance from the prediction point increases, revealing its limitations in capturing long-term dependencies. Finally, the post-hoc explainable models, MTEX and XCM, reveal notable discrepancies compared to the existing literature. Both models fail to align with domain knowledge in terms of VI and TI, highlighting the inherent limitations of post-hoc explainability methods in accurately capturing complex multivariate time-series dynamics.

V. ABLATION STUDIES

We conducted a series of ablation and analytical experiments to comprehensively evaluate DLFormer from the perspectives of architectural effectiveness, scalability, computational efficiency, and interpretability. Specifically, this section examines the contributions of DL Embedding components, the role of attention operations in the encoder-decoder architecture, the effect of model-capacity scaling, and the explanatory behavior of TVA attention maps. We address the following five research questions:

- Are the components of DL Embedding effective? (A)
- Is the inclusion of attention layers in DL Encoder and DL Decoder architectures justified? (B)
- How does increasing the model capacity of DLFormer affect forecasting performance and computational cost? (C)
- Does the TVA attention map effectively capture the temporal periodicity associated with time lags? (D)
- Can the TVA attention maps reasonably explain a single sample under specific conditions? (E)

TABLE XIV
DL EMBEDDING COMPONENTS EFFECT ANALYSIS RESULTS

Case	GSPE+LSPE		w/o GSPE		w/o LSPE	
Metrics	MSE	MAE	MSE	MAE	MSE	MAE
PAQ	**0.876**	**0.642**	0.966	0.682	0.925	0.663
SML	**0.187**	**0.251**	0.190	0.257	0.193	0.255
Volatility	**0.383**	**0.374**	0.484	0.411	0.443	0.393
PM2.5	**0.504**	**0.490**	0.51	0.495	0.506	0.491
Exchange	**0.096**	**0.196**	0.111	0.203	0.098	0.197
Weather	**0.160**	**0.242**	0.19	0.277	0.162	0.256
ETTh1	**0.089**	**0.212**	0.112	0.234	0.095	0.218
ETTm1	**0.067**	**0.165**	0.068	0.166	0.069	0.168
ETTh2	**0.112**	**0.237**	0.115	0.242	0.114	0.24
ETTm2	**0.068**	**0.161**	0.09	0.183	0.078	0.172

TABLE XV
ANALYSIS RESULTS OF ATTENTION OPERATION WITHIN DL ENCODER AND DL DECODER

Enc / Dec	AL / AL		LL / AL		AL / LL		LL / LL	
Metrics	MSE	MAE	MSE	MAE	MSE	MAE	MSE	MAE
PAQ	**0.876**	**0.642**	0.877	0.647	0.896	0.648	0.948	0.671
SML	**0.187**	**0.251**	0.196	0.256	0.196	0.263	0.192	0.26
Volatility	**0.383**	**0.374**	0.478	0.404	0.552	0.445	0.387	0.382
PM2.5	**0.504**	**0.490**	0.507	0.497	0.506	0.493	0.509	0.499
Exchange	**0.096**	**0.196**	0.098	0.197	0.116	0.209	0.098	0.198
Weather	**0.160**	**0.242**	**0.160**	0.248	0.218	0.290	0.165	0.255
ETTh1	0.089	0.212	0.091	0.215	**0.087**	0.213	0.086	**0.208**
ETTm1	**0.067**	**0.165**	0.071	0.171	0.068	0.166	0.069	0.169
ETTh2	**0.112**	**0.237**	0.12	0.248	0.117	0.242	0.126	0.251
ETTm2	**0.068**	**0.161**	0.108	0.195	0.074	0.167	0.082	0.176

A. Analysis of the Effects of DL Embedding Components

To analyze the contribution of each component within DL Embedding, we conducted ablation experiments by selectively removing combinations of GSPE and LSPE. We first

TABLE XVI
PERFORMANCE COMPARISON BETWEEN LARGE DLFORMER AND TIMELLM WITH LLM-LEVEL HIDDEN DIMENSIONS

Dataset	PAQ			SML			Volatility			PM2.5			Exchange		
Metrics	MSE	MAE	COR	MSE	MAE	COR	MSE	MAE	COR	MSE	MAE	COR	MSE	MAE	COR
DLFormer	1.100	0.740	0.612	0.405	0.490	0.818	0.801	0.629	**0.348**	0.718	0.633	0.156	0.182	0.318	0.38
Large DLFormer	**1.079**	**0.722**	**0.645**	**0.368**	**0.462**	**0.845**	**0.716**	**0.605**	0.342	**0.704**	**0.631**	**0.161**	**0.130**	**0.279**	**0.43**
TimeLLM	1.257	0.826	0.532	0.417	0.517	**0.845**	0.749	0.636	0.105	0.908	0.688	0.131	0.294	0.381	-0.02

TABLE XVII
PERFORMANCE COMPARISON BETWEEN LARGE DLFORMER AND TIMELLM WITH LLM-LEVEL HIDDEN DIMENSIONS

Dataset	Weather			ETTh1			ETTm1			ETTh2			ETTm2		
Metrics	MSE	MAE	COR	MSE	MAE	COR	MSE	MAE	COR	MSE	MAE	COR	MSE	MAE	COR
DLFormer	0.293	0.396	0.430	0.149	0.300	0.198	0.124	0.263	0.218	**0.182**	0.337	0.584	0.121	0.261	0.657
Large DLFormer	0.277	0.383	**0.432**	0.134	0.288	0.213	0.099	0.238	**0.240**	0.183	**0.335**	**0.608**	**0.117**	**0.252**	**0.676**
TimeLLM	**0.227**	**0.340**	0.320	**0.086**	**0.226**	**0.214**	**0.054**	**0.175**	0.191	0.201	0.352	0.607	0.135	0.274	0.607

TABLE XVIII
COMPUTATIONAL COST COMPARISON OF DLFORMER, LARGE DLFORMER, AND TIMELLM

Models	d_{model}	Params	MACs	GPU Memory
DLFormer	128	0.55M	5.31G	0.73GB
Large DLFormer	1024	16.64M	179.16G	0.99GB
TimeLLM	1024	50.88M	40.81G	6.95GB

TABLE XIX
THEORETICAL ATTENTION COMPLEXITY UNDER DIFFERENT TOKENIZATION STRATEGIES

Tokenization strategy	Token length	Attention computation	Attention memory
Time-token attention	T	$O(T^2 d_{model})$	$O(T^2)$
Variable-token attention	F	$O(F^2 d_{model})$	$O(F^2)$
Patch-token attention	N_p	$O({N_p}^2 d_{model})$	$O({N_p}^2)$
Time-variable token attention (Ours)	$T \times F$	$O((TF)^2 d_{model})$	$O((TF)^2)$

performed forecasting for all prediction lengths (1, 3, 6, 12, 96, 192, 336, and 720) and then calculated the average performance across all forecasting horizons. Table XIII presents the results of the component analysis of DL Embedding.

The values represent the average prediction results obtained from five independent experimental runs. The ablation study shows that DLFormer with both GSPE and LSPE achieves the best performance, supporting the usefulness of positional decomposition in DL Embedding for MTSF tasks. In particular, the performance degradation observed when either GSPE or LSPE is removed suggests that the model's ability to jointly preserve within-variable temporal order and global time-variable identity may be weakened.

B. Analysis of Attention Operation within DL Encoder and DL Decoder

We conducted an experiment to compare the performance obtained by replacing the attention layers in the original DL Encoder and DL Decoder with linear layers. Table XIV presents the results of the attention operation. The values reflect forecasts for all prediction lengths (1, 3, 6, 12, 96, 192, 336, and 720) and the average performance across these horizons. These results confirm that the attention layers effectively capture the relationships in the input data when combined with DL Embeddings in the encoder and decoder networks. In contrast, the encoder-decoder structure with linear layers fails to effectively learn the interactions between temporal and variable dimensions. The weight-sharing nature of linear layers across multiple dimensions likely limits their ability to model complex cross-dimensional relationships when combined with DL Embeddings.

C. Scalability and Computational Cost Analysis of DLFormer

Recent MTSF studies have reported that large-scale models, such as TimeLLM [56], can achieve strong predictive performance by leveraging increased model capacity and pretrained language-model representations. Motivated by this trend, we further analyzed the scalability of DLFormer by increasing its hidden dimension from 128 to 768. This extended model is referred to as Large DLFormer. Tables XVI and XVII present the forecasting performance comparison between Large DLFormer and TimeLLM, while Table XVIII summarizes their computational costs.

The results show that hidden-dimension scaling improves the average predictive performance of DLFormer by 8.98%, suggesting that the proposed DL Embedding and TVAL mechanisms benefit from increased model capacity. This performance improvement indicates that a larger hidden

dimension can strengthen the representation of flattened time-variable tokens and enhance the ability to model variable-specific lag dependencies. However, this gain is accompanied by increased computational cost. Large DLFormer has substantially fewer parameters and lower GPU memory usage than TimeLLM, but requires higher MACs because it performs dense attention over variable-lag tokens. This shows that the scalability of DLFormer involves a trade-off between explicit time-variable representation and operation-level computational cost.

To further clarify this scalability characteristic, Table XIX presents the theoretical attention complexity according to different tokenization strategies. Unlike time-token or variable-token Transformer variants, DLFormer attends over a flattened sequence of length $T \times F$, resulting in an attention computational complexity of $O((TF)^2 d_{model})$ and an attention memory complexity of $O((TF)^2)$. This limitation suggests the need for future extensions using sparse attention, patch-based tokenization, or variable selection mechanisms to improve scalability while preserving the interpretability benefits of TVAL.

D. Analysis of the Time Lag Capturing Capability of the Time-Variable Aware Attention Map

TVA attention map serve as a key interpretability component in DLFormer, providing insights into the learned temporal dependencies for each prediction step. Specifically, it visualizes the attention weights across all past variables and time steps, contributing to each future prediction point. Similarly, the time-lagged cross-correlation (TLCC) is a statistical tool designed to quantify the temporal correlation between past variables and future prediction targets [57]. High TLCC values indicate that a specific time point of the given variable exhibits a strong correlation with the future horizon. Given this, if DLFormer effectively learns the intended variable-time relationships through DL Embedding, we hypothesize that the TVA attention map should exhibit patterns comparable to those observed in TLCC. Fig. 8 compares TVA attention maps and TLCC across multiple datasets. Subfigures (a)–(j) correspond to the PAQ, SML, Volatility, PM2.5, Exchange, Weather, ETTh1, ETTm1, ETTh2, and ETTm2 datasets, respectively. All comparisons were performed in the univariate prediction setting. For each dataset, the left panel displays the learned TVA attention map, while the right panel presents the corresponding TLCC heatmap for the same dataset. TLCC is visualized using the same axis configuration as the attention map, where the x-axis represents the time steps for each variable (with the rightmost position indicating the autoregressive target variable), and the y-axis denotes the forecast time step. The values reflect the average correlation across all samples for each (variable, time) pair. As expected in supervised forecasting tasks, the TVA attention map predominantly assigns high attention weights to past time steps of the target variable, indicating their strong predictive influence. Beyond this autoregressive effect, a

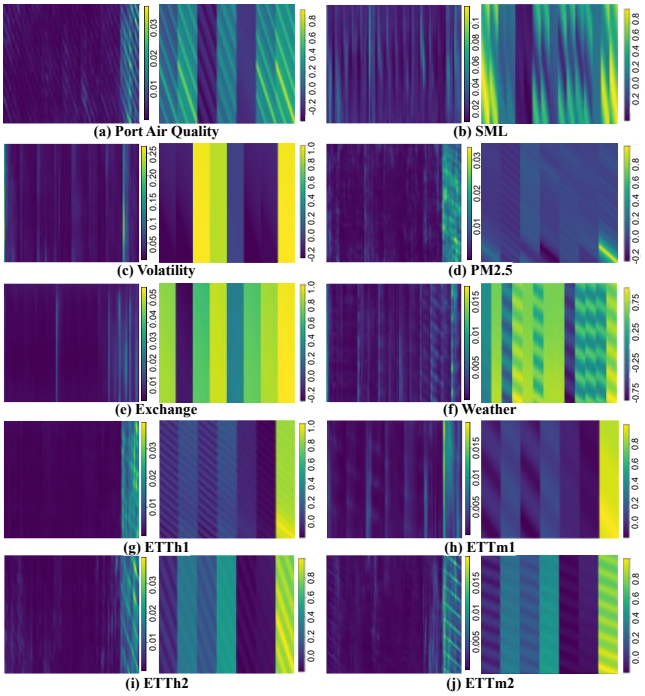

Fig. 8. Visualization of the comparison results between the time-variable-aware attention map and time lagged cross-correlation map.

notable observation emerges: for most datasets, except those with inherently stochastic characteristics such as exchange and volatility, the TVA attention map reveals patterns that closely align with the high-correlation regions in TLCC.

Moreover, similar to TLCC, the TVA attention map exhibits visually distinct grid-like structures that demarcate the temporal influence of each variable. This confirms that the DL Embedding framework effectively captures the variable-specific temporal relationships as intended. While prior studies [58], [59] have suggested that token embeddings in Transformers can degenerate into random noise, our results demonstrate otherwise. The comparison indicates that DLFormer, through its specialized DL Embedding and attention mechanisms, learns semantically meaningful representations that reflect variable-specific time dependencies critical for MTSF.

E. Local Explainability Analysis of the Time-Variable Aware Attention Map

To evaluate the local explainability of the TVA attention map, we conducted experiments using two datasets: (a) SML and (b) PM2.5, both characterized by distinct and observable changes under specific periods or conditions. The goal is to assess whether the TVA attention map can effectively highlight meaningful variations. Fig. 9 presents the local explainability analysis results for both datasets. The top row visualizes time-series plots for periods of significant change. The bottom-left quadrant illustrates the evolution of input signals across sequential samples and the corresponding TVA attention maps extracted by DLFormer. The bottom-right quadrant visualizes

This article has been accepted for publication in IEEE Access. This is the author's version which has not been fully edited and content may change prior to final publication. Citation information: DOI 10.1109/ACCESS.2026.3713022

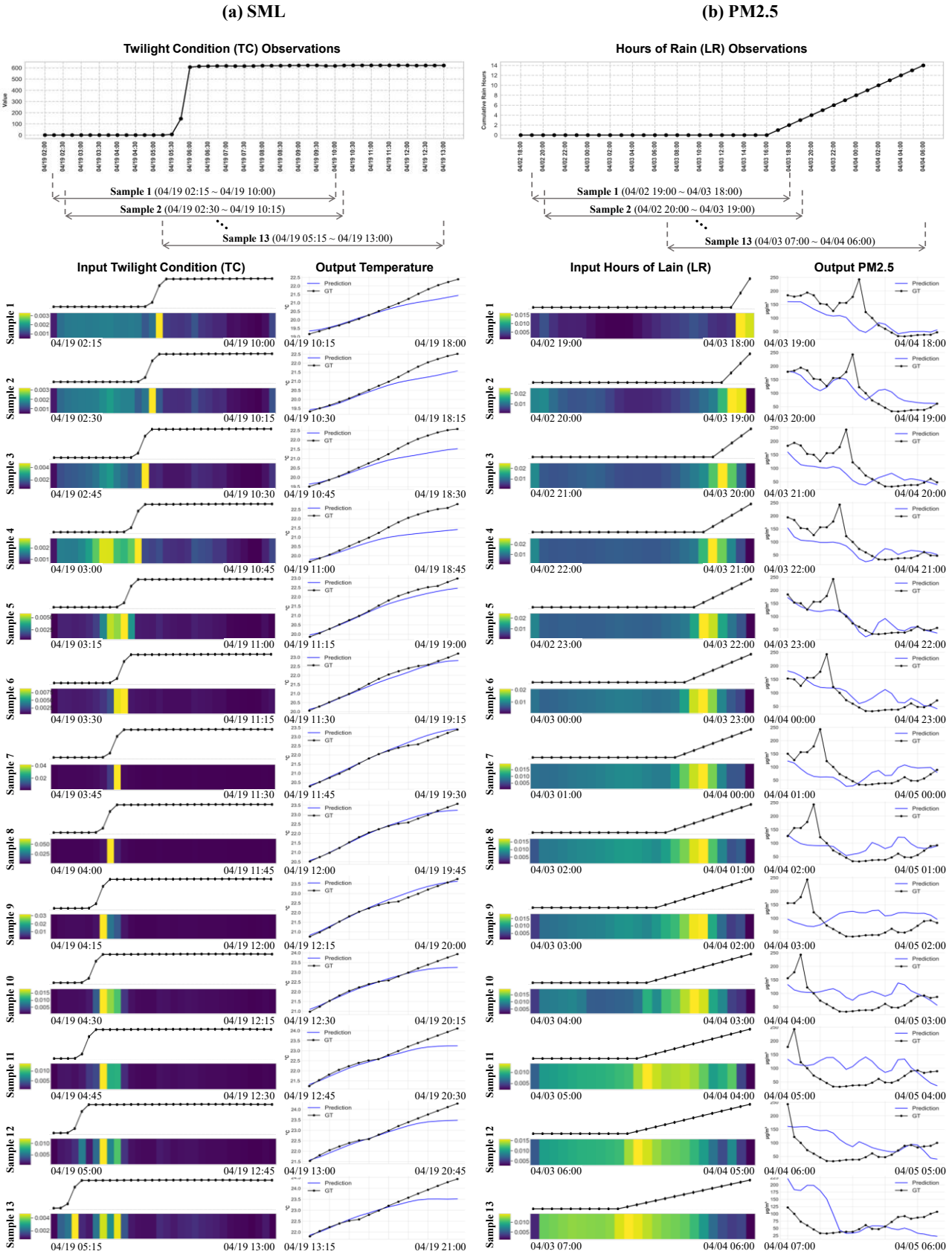

Fig. 9. Detailed visualization of the time-variable-aware attention map for local explainability in (a) SML, (b) PM2.5 data.

the predicted values of the target variable over the forecast horizon. All results are derived from samples drawn from the test set.

a) SML Dataset: In the SML dataset, the input length is set to 32 steps (8 h) and the prediction length to 16 steps (4 h). The input variable used for analysis is the Twilight Condition (TC). The top graph shows TC changing sharply between 5:00 a.m. and 6:00 a.m., aligning with sunrise. In Sample 1, the TVA attention map assigns high weights to the time step where TC begins to rise, precisely identifying the critical point. Subsequent samples consistently focus on this sunrise transition, correlating with a gradual increase in the predicted indoor temperature. This indicates that DLFormer successfully captures the temperature rise following sunrise, providing a meaningful and interpretable local explanation.

b) PM2.5 Dataset: In the PM2.5 dataset, the input and prediction lengths are set to 24 steps (24 h). The input variable is Hours of Rain (LR). The time-series plot indicates that rainfall begins around 16:00 on April 3, with the cumulative LR value increasing linearly thereafter. In Sample 1, the TVA attention map accurately focuses on the moment rain begins. Consequently, the model predicts a decrease in PM2.5 concentration, consistent with established domain knowledge [52]-[54]. This pattern continues in subsequent samples, where the attention focuses on the onset and persistence of rainfall, leading to flattened, lower PM2.5 predictions that reflect the expected atmospheric cleansing effect. These findings suggest that DLFormer captures the impact of specific input variables and provides locally interpretable explanations for its forecasts.

These results confirm that the TVA attention map demonstrates strong local explainability, accurately identifying variable-specific temporal effects under specific conditions and producing interpretable attention behaviors aligned with real-world dynamics.

VI. DISCUSSION

The results of this study suggest that DLFormer can provide structural explainability by redefining the token space of multivariate time-series forecasting in a flattening-based form. This design aligns the unit of modeling with the unit of explanation, which is particularly important in MTSF tasks where both variable identity and temporal lag determine predictive relevance. Within this structure, the experimental results support the contribution of LSPE and GSPE to preserving token identity. Accordingly, the proposed TVAL strategy should not be viewed merely as an input-reshaping operation, but rather as an inductive bias that simultaneously determines the modeling unit of each token and the granularity of the resulting explanation. This perspective further suggests that DLFormer-style tokenization may be extended to more structured domains, including time-variable-space, time-variable-node, and time-variable-frequency representations. In such settings, additional structural axes could be incorporated into interpretable forecasting tokens.

The scalability analysis further shows that DLFormer can benefit from increased model capacity, although this improvement entails a trade-off between predictive performance and computational efficiency. Owing to the TVAL strategy, the computational cost of DLFormer is more sensitive to both input length and variable dimensionality than that of conventional Transformer architectures. Nevertheless, increasing the hidden dimension enhances the model's capacity to represent complex interactions among variable-lag tokens, thereby improving forecasting performance. Therefore, the scalability of DLFormer should not be interpreted simply as a "larger is better" effect, but rather as evidence that increased capacity can be beneficial when applied to a structurally meaningful time-variable representation. The current analysis, however, focuses primarily on model-capacity scaling, and additional investigation is required from the perspective of dataset scalability. Empirical evaluation of predictive performance and computational cost on very high-dimensional datasets containing hundreds or thousands of variables remains an important direction for future research. In such large-scale model and data settings, efficient designs such as sparse attention, patch-based tokenization, factorized attention, or variable selection may be required to reduce the computational burden while preserving the advantages of TVAL.

Finally, DLFormer provides a useful perspective on attention-based explainability. A common limitation of attention-based XAI is that attention weights are not always equivalent to explanations, particularly when the underlying tokens lack clear semantic meaning. DLFormer partially alleviates this issue by defining attention tokens at a fine-grained variable-lag level, thereby allowing the attention map to be interpreted more directly as a temporal-variable importance pattern. Nevertheless, these attention maps do not fully explain all operating principles of the model and should be understood as indicators of predictive relevance rather than direct causal effects. For this reason, the attention analysis was complemented by both quantitative and qualitative evaluations, including robustness analysis, perturbation-based faithfulness evaluation, and domain-oriented case studies. These evaluations support the practical validity of the learned explanations. Nevertheless, establishing rigorous statistical evaluation protocols for explanation quality remains a promising direction for future research, particularly for quantitatively comparing explanation behaviors and quality across different models and datasets.

VII. CONCLUSION

In this study, we proposed DLFormer, a Transformer-based framework for explainable multivariate time-series forecasting. DLFormer introduced Time-Variable-Aware Learning (TVAL) and Distributed Lag (DL) Embedding to represent variable-specific temporal lags as token-level modeling units. This design aligned the unit of forecasting with the unit of explanation and enabled the model to capture

both local within-variable dependencies and global cross-variable lagged interactions. Comprehensive experiments on ten benchmark and real-world datasets showed that DLFormer improved forecasting performance while providing robust, faithful, and valid explanations from temporal and variable perspectives. Through these contributions, DLFormer provided a structured approach for improving both forecasting performance and explainability in high-dimensional MTSF tasks.

Despite these findings, DLFormer has several limitations that should be addressed in future work. First, because TVAL represents multivariate time-series inputs as variable-lag tokens, its computational and memory costs can increase substantially in very high-dimensional settings or with long input windows. Future work will therefore aim to improve scalability through sparse attention, patch-based variable-lag tokenization, factorized attention, or variable selection mechanisms. Second, although the TVA attention map provides structured predictive importance signals, it should not be interpreted as a strict causal explanation, and its validity still depends on the assumption that attention over variable-lag tokens reflects input relevance. To strengthen the reliability of the learned explanations, future studies will combine TVA attention maps with intervention-based, counterfactual, or causal validation frameworks. In addition, DLFormer will be extended to irregular and missing time-series settings by incorporating time-gap-aware embeddings and masking-aware attention mechanisms, making it more scalable, robust, and actionable for real-world forecasting applications.

REFERENCES

[1] S. Sim, D. Kim, and S. C. Jeong, "Temporal attention gate network with temporal decomposition for improved prediction accuracy of univariate time-series data," in *Proc. Int. Conf. Artif. Intell. Inf. Commun*. (ICAIIC), 2023, pp. 122–127.

[2] J. Lee, S. Lee, Y. Kim, D. Kim, and S. Sim, "FEATHer: Fourier-efficient adaptive temporal hierarchy forecaster for time-series forecasting," *arXiv preprint arXiv:2601.11350*, 2026.

[3] Y. Kim, J. Lee, H. Bae, and S. Sim, "How will Arctic shipping emissions evolve? A spatiotemporal topology-aware transformer approach," *Transp. Res. Part D, Transp. Environ., vol. 151*, Art. no. 105134, 2026.

[4] S. Sim, J. H. Park, and H. Bae, "Deep collaborative learning model for port-air pollutants prediction using automatic identification system," *Transp. Res. Part D, Transp. Environ*., vol. 111, Art. no. 103431, 2022.

[5] C. An, Z. Lu, Y. Ma, W. Zhao, and X. Chen, "Prediction of multivariate spatial-temporal series data based on adaptive spatial-temporal information," *IEEE Trans. Big Data*, early access, 2025, doi: 10.1109/TBDATA.2025.3588086.

[6] J. Lee, D. Kim, and S. Sim, "Temporal multi-features representation learning-based clustering for time-series data," *IEEE Access*, vol. 12, pp. 87675–87690, 2024.

[7] S. Sim, D. Kim, and H. Bae, "Correlation recurrent units: A novel neural architecture for improving the predictive performance of time-series data," *IEEE Trans. Pattern Anal. Mach. Intell*., vol. 45, no. 12, pp. 14266–14283, Dec. 2023.

[8] S. Wang, Y. Zhang, X. Lin, Y. Hu, Q. Huang, and B. Yin, "Dynamic hypergraph structure learning for multivariate time series forecasting," *IEEE Trans. Big Data*, vol. 10, no. 4, pp. 556–567, Jul.–Aug. 2024.

[9] K. Wang, C. Xu, Y. Zhang, S. Guo, and A. Y. Zomaya, "Robust big data analytics for electricity price forecasting in the smart grid," *IEEE Trans. Big Data*, vol. 5, no. 1, pp. 34–45, Mar. 2019.

[10] P. Zhang, Y. Jia, J. Gao, W. Song, and H. Leung, "Short-term rainfall forecasting using multi-layer perceptron," *IEEE Trans. Big Data*, vol. 6, no. 1, pp. 93–106, Mar. 2020.

[11] C. O. Retzlaff, A. Angerschmid, A. Saranti, D. Schneeberger, R. Roettger, H. Mueller, and A. Holzinger, "Post-hoc vs. ante-hoc explanations: xAI design guidelines for data scientists," *Cogn. Syst. Res*., vol. 86, Art. no. 101243, 2024.

[12] A. Theissler, F. Spinnato, U. Schlegel, and R. Guidotti, "Explainable AI for time series classification: A review, taxonomy and research directions," *IEEE Access*, vol. 10, pp. 100700–100724, 2022.

[13] X.-H. Li, C. C. Cao, Y. Shi, W. Bai, H. Gao, L. Qiu, C. Wang, Y. Gao, S. Zhang, X. Xue, and L. Chen, "A survey of data-driven and knowledge-aware explainable AI," *IEEE Trans. Knowl. Data Eng*., vol. 34, no. 1, pp. 29–49, Jan. 2022.

[14] B. Crook, M. Schlüter, and T. Speith, "Revisiting the performance-explainability trade-off in explainable artificial intelligence (XAI)," in *Proc. IEEE Int. Requirements Eng. Conf. Workshops* (REW), 2023, pp. 316–324.

[15] S. Almon, "The distributed lag between capital appropriations and expenditures," *Econometrica*, vol. 33, no. 1, pp. 178–196, Jan. 1965.

[16] Z. Lyu, Y. Wu, J. Lai, M. Yang, C. Li, and W. Zhou, "Knowledge enhanced graph neural networks for explainable recommendation," *IEEE Trans. Knowl. Data Eng*., vol. 35, no. 5, pp. 4954–4968, May 2023.

[17] L. Hu, X. Wang, Y. Liu, N. Liu, M. Huai, L. Sun, and D. Wang, "Towards stable and explainable attention mechanisms," *IEEE Trans. Knowl. Data Eng*., early access, 2025.

[18] L. Longo, R. Goebel, F. Lecue, P. Kieseberg, and A. Holzinger, "Explainable artificial intelligence: Concepts, applications, research challenges and visions," in *Proc. Int. Cross-Domain Conf. Mach. Learn. Knowl. Extract*. (CD-MAKE). Cham, Switzerland: Springer, 2020, pp. 1–16.

[19] R. Saleem, B. Yuan, F. Kurugollu, A. Anjum, and L. Liu, "Explaining deep neural networks: A survey on the global interpretation methods," *Neurocomputing*, vol. 513, pp. 165–180, 2022.

[20] A. L. Alfeo, A. G. Zippo, V. Catrambone, M. G. Cimino, N. Toschi, and G. Valenza, "From local counterfactuals to global feature importance: Efficient, robust, and model-agnostic explanations for brain connectivity networks," *Comput. Methods Programs Biomed*., vol. 236, Art. no. 107550, 2023.

[21] M. Setzu, R. Guidotti, A. Monreale, F. Turini, D. Pedreschi, and F. Giannotti, "GlocalX: From local to global explanations of black box AI models," *Artif. Intell*., vol. 294, Art. no. 103457, 2021.

[22] S. M. Lundberg and S.-I. Lee, "A unified approach to interpreting model predictions," in *Proc. Adv. Neural Inf. Process. Syst*. (NeurIPS), vol. 30, 2017.

[23] M. T. Ribeiro, S. Singh, and C. Guestrin, "Why should I trust you? Explaining the predictions of any classifier," in *Proc. 22nd ACM SIGKDD Int. Conf. Knowl. Discovery Data Mining*, 2016, pp. 1135–1144.

[24] R. R. Selvaraju, M. Cogswell, A. Das, R. Vedantam, D. Parikh, and D. Batra, "Grad-CAM: Visual explanations from deep networks via gradient-based localization," in *Proc. IEEE Int. Conf. Comput. Vis*. (ICCV), 2017, pp. 618–626.

[25] A. Chattopadhay, A. Sarkar, P. Howlader, and V. N. Balasubramanian, "Grad-CAM++: Generalized gradient-based visual explanations for deep convolutional networks," in *Proc. IEEE Winter Conf. Appl. Comput. Vis*. (WACV), 2018, pp. 839–847.

[26] Y. Wang, "A comparative analysis of model agnostic techniques for explainable artificial intelligence," *Res. Rep. Comput. Sci*., vol. 3, no. 2, pp. 25–33, 2024.

[27] L. Tronchin, E. Cordelli, L. R. Celsi, D. Maccagnola, M. Natale, P. Soda, and R. Sicilia, "Translating image XAI to multivariate time series," *IEEE Access*, vol. 12, pp. 27484–27500, 2024.

[28] B. Lim, S. Ö. Arık, N. Loeff, and T. Pfister, "Temporal fusion transformers for interpretable multi-horizon time series forecasting," *Int. J. Forecast*., vol. 37, no. 4, pp. 1748–1764, 2021.

[29] L. Pantiskas, K. Verstoep, and H. Bal, "Interpretable multivariate time series forecasting with temporal attention convolutional neural networks," in *Proc. IEEE Symp. Ser. Comput. Intell*. (SSCI), 2020, pp. 1687–1694.

[30] K. Fauvel, T. Lin, V. Masson, É. Fromont, and A. Termier, "XCM: An explainable convolutional neural network for multivariate time series classification," *Mathematics*, vol. 9, no. 23, Art. no. 3137, 2021.
[31] J. Droste, R. Fuchs, H. Deters, J. Klünder, and K. Schneider, "Explainability requirements for time series forecasts: A study in the energy domain," in *Proc. IEEE Int. Requirements Eng. Conf*. (RE), 2024, pp. 229–239.
[32] Y. Liu, T. Hu, H. Zhang, H. Wu, S. Wang, L. Ma, and M. Long, "iTransformer: Inverted transformers are effective for time series forecasting," in *Proc. Int. Conf. Learn. Represent*. (ICLR), 2024.
[33] A. Zeng, M. Chen, L. Zhang, and Q. Xu, "Are transformers effective for time series forecasting?" in *Proc. AAAI Conf. Artif. Intell*., vol. 37, no. 9, 2023, pp. 11121–11128.
[34] Y. Nie, N. H. Nguyen, P. Sinthong, and J. Kalagnanam, "A time series is worth 64 words: Long-term forecasting with transformers," in *Proc. Int. Conf. Learn. Represent*. (ICLR), 2023.
[35] H. Wu, J. Xu, J. Wang, and M. Long, "Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting," in *Proc. Adv. Neural Inf. Process. Syst*. (NeurIPS), vol. 34, 2021, pp. 22419–22430.
[36] H. Wu, T. Hu, Y. Liu, H. Zhou, J. Wang, and M. Long, "TimesNet: Temporal 2D-variation modeling for general time series analysis," in *Proc. Int. Conf. Learn. Represent*. (ICLR), 2023.
[37] T. Guo, T. Lin, and N. Antulov-Fantulin, "Exploring interpretable LSTM neural networks over multi-variable data," in Proc. 36th Int. Conf. Mach. Learn. (ICML), *Proc. Mach. Learn. Res*., vol. 97, 2019, pp. 2494–2504.
[38] H. Jang, C. Kim, and E. Yang, "TIMING: Temporality-aware integrated gradients for time series explanation," *arXiv preprint arXiv:2506.05035*, 2025.
[39] J. Bento, P. Saleiro, A. F. Cruz, M. A. T. Figueiredo, and P. Bizarro, "TimeSHAP: Explaining recurrent models through sequence perturbations," in *Proc. 27th ACM SIGKDD Conf. Knowl. Discovery Data Mining* (KDD), 2021, pp. 2565–2573.
[40] A. M. Calero and A. Rasheed, "ChronoSHAP: Revealing temporal patterns from transformers and LTSF-linear models in time series forecasting," *Knowl.-Based Syst*., vol. 331, Art. no. 114802, 2026.
[41] Z. Wang, I. Miliou, I. Samsten, and P. Papapetrou, "Counterfactual explanations for time series forecasting," in *Proc. IEEE Int. Conf. Data Mining* (ICDM), 2023, pp. 1391–1396.
[42] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, "Attention is all you need," in *Proc. Adv. Neural Inf. Process. Syst*. (NeurIPS), vol. 30, 2017.
[43] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," *arXiv preprint arXiv:1412.6980*, 2014.
[44] D. Alvarez-Melis and T. S. Jaakkola, "Towards robust interpretability with self-explaining neural networks," in *Proc. Adv. Neural Inf. Process. Syst*. (NeurIPS), vol. 31, 2018, pp. 7786–7795.
[45] S. Hooker, D. Erhan, P.-J. Kindermans, and B. Kim, "A benchmark for interpretability methods in deep neural networks," in *Proc. Adv. Neural Inf. Process. Syst*. (NeurIPS), vol. 32, 2019, pp. 9737–9748.
[46] Y. Qin, D. Song, H. Chen, W. Cheng, G. Jiang, and G. W. Cottrell, "A dual-stage attention-based recurrent neural network for time series prediction," in *Proc. 26th Int. Joint Conf. Artif. Intell*. (IJCAI), 2017, pp. 2627–2633.
[47] J. Herzen, F. Lässig, S. G. Piazzetta, T. Neuer, L. Tafti, G. Raille, and G. Grosch, "Darts: User-friendly modern machine learning for time series," *J. Mach. Learn. Res*., vol. 23, no. 124, pp. 1–6, 2022.
[48] G. P. Gobbi, L. Di Liberto, and F. Barnaba, "Impact of port emissions on EU-regulated and non-regulated air quality indicators: The case of Civitavecchia (Italy)," *Sci. Total Environ*., vol. 719, Art. no. 134984, 2020.
[49] J. Pan, Y. Wang, X. Qin, N. K. Gali, Q. Fu, and Z. Ning, "Spatiotemporal analysis of complex emission dynamics in port areas using high-density air sensor network," *Toxics*, vol. 12, no. 10, Art. no. 760, 2024.
[50] Z. Song et al., "Significant reductions of urban daytime ozone by extremely high concentration NOx from ship's emissions: A case study," *Atmos. Pollut. Res*., vol. 15, no. 7, Art. no. 102142, 2024.
[51] D. M. Stansel, N. M. Laurendeau, and D. W. Senser, "CO and NOx emissions from a controlled-air burner: Experimental measurements and exhaust correlations," *Combust. Sci. Technol*., vol. 104, no. 4–6, pp. 207–234, 1995.
[52] S. Paz, P. Goldstein, L. Kordova-Biezuner, and L. Adler, "Benzene patterns in different urban environments and a prediction model for benzene rates based on NOx values," in *Proc. EGU Gen. Assem. Conf. Abstr*., 2017, p. 4750.
[53] B. Chen, S. Lu, S. Li, and B. Wang, "Impact of fine particulate fluctuation and other variables on Beijing's air quality index," *Environ. Sci. Pollut. Res*., vol. 22, pp. 5139–5151, 2015.
[54] G. Peng, A. B. Umarova, and G. S. Bykova, "The temporal and spatial changes of Beijing's PM2.5 concentration and its relationship with meteorological factors from 2015 to 2020," *Geogr. Environ. Sustain*., vol. 14, no. 3, pp. 73–81, 2021.
[55] Y. Shi et al., "Analysis of characteristics of atmosphere particulate matter pollution in Beijing during the fall and winter of 2012 to 2013," *Ecol. Environ. Sci*., vol. 22, pp. 1571–1577, 2013.
[56] M. Jin et al., "Time-LLM: Time series forecasting by reprogramming large language models," in *Proc. Int. Conf. Learn. Represent*. (ICLR), 2024.
[57] P. Xu, B.-S. Wang, W. Zhang, J.-M. Lin, and Y. Chen, "Estimating seismic attenuation using cross-correlation function," *Chin. J. Geophys*., vol. 49, no. 6, pp. 1738–1744, 2006.
[58] S. Jain and B. C. Wallace, "Attention is not explanation," *arXiv preprint arXiv:1902.10186*, 2019.
[59] A. Bibal, R. Cardon, D. Alfter, R. Wilkens, X. Wang, T. François, and P. Watrin, "Is attention explanation? An introduction to the debate," in *Proc. Annu. Meet. Assoc. Comput. Linguist*. (ACL), 2022, pp. 3889–3900.

Younghwi Kim received his B.S. in Creative Software Engineering from Dong-eui University, Republic of Korea, in 2024, where he is currently pursuing an M.S. degree in the Graduate School of AI Convergence Engineering at Changwon National University, Changwon, Republic of Korea. His research interests include data mining and deep learning, with a particular focus on time-series neural network structure design and explainable artificial intelligence.

Dohee Kim (Member, IEEE) received her B.S. in Industrial Engineering from Pusan National University, Republic of Korea, in 2019 and her M.S. and Ph.D. in Industrial Engineering from Pusan National University, Republic of Korea, in 2024. She is currently an assistant professor with the Department of Artificial Intelligence Convergence Engineering at Changwon National University, Republic of Korea. Her research interests include deep learning, optimization, representation learning, and data mining.

Joongrok Kim received his M.S. in Graduate Program in Biometrics and his Ph.D. degree in Electrical and Electronic Engineering from Yonsei University, Republic of Korea, in 2008 and 2014, respectively. From 2014 to 2024, he was a researcher at the LG Electronics AI Lab, where he worked on applying artificial intelligence technologies to

various domains including automotive and home appliances. He is currently an associate professor with the Department of Artificial Intelligence Convergence Engineering at Changwon National University, Republic of Korea. His research interests include computer vision, deep learning, time series forecasting, and AI applications in industrial systems.

Sunghyun Sim (Member, IEEE) received his B.S. in Statistics from Pusan National University, Republic of Korea, in 2016 and his Ph.D. in Industrial Engineering from Pusan National University, Republic of Korea, in 2021. He is currently an assistant professor with the Department of Artificial Intelligence Convergence Engineering at Changwon National University, Republic of Korea. His research interests include data mining, process mining, and deep learning, with a particular focus on neural layer structure design, lightweight deep learning models, and explainable artificial intelligence.