Journal of Computer and Communications, 2026, 14(3), 174-195
https://www.scirp.org/journal/jcc
ISSN Online: 2327-5227
ISSN Print: 2327-5219

KAN-Transformer Fusion with Mixture of Experts for Temporal Imputation of Spatiotemporal Air Pollution Data

Jiawen Ding

School of Mathematics and Statistics, Guilin University of Technology, Guilin, China
Email: JiawenD@outlook.com

How to cite this paper: Ding, J.W. (2026) KAN-Transformer Fusion with Mixture of Experts for Temporal Imputation of Spatiotemporal Air Pollution Data. *Journal of Computer and Communications*, **14**, 174-195.
https://doi.org/10.4236/jcc.2026.143009

Received: February 28, 2026
Accepted: March 23, 2026
Published: March 26, 2026

Abstract

Air pollution has become a pressing challenge to global public health and environmental governance. Accurate analysis of pollutant concentrations critically depends on the completeness of monitoring data; however, the widespread presence of missing values in real-world datasets significantly compromises both assessment reliability and predictive performance. To address this issue, this study proposes a three-dimensional time-feature-space model, which integrates a Kolmogorov-Arnold Network (KAN) with a Transformer architecture, referred to as KT. The KAN module, equipped with an expert mixing mechanism, captures complex nonlinear temporal dynamics of pollutants, which are subsequently processed by the Transformer to model interdependencies among multiple pollutants via self-attention. A spatial feature selection strategy based on Spearman correlation is further employed to extract key spatiotemporal interactions through channel mixing or independent dynamic processing. Empirical evaluation on air quality monitoring data collected in Beijing from January 2023 to October 2024 shows that the proposed model reduces the mean absolute error (MAE) by 1.1% - 17.1% compared with several SOTA benchmark methods. These results clearly demonstrate the robustness and effectiveness of the proposed approach in estimating complex and incomplete air pollution data.

Keywords

Air Quality, Missing Values Imputation, Kolmogorov-Arnold Network (KAN), Transformer, Deep Learning

1. Introduction

With the acceleration of global industrialization and urbanization, air pollution has become an escalating environmental challenge requiring urgent attention. In recent years, concentrations of key air pollutants—such as fine particulate matter ($PM_{2.5}$), sulphur dioxide (SO_2), and nitrogen oxides (NO_x)—have shown upward trends in several urban areas across China. Multiple monitoring datasets indicate that, in certain regions, pollutant levels have exceeded the air quality guideline values recommended by the World Health Organization (WHO) [1]. This severe air pollution poses a direct threat to public health, increasing the incidence of chronic conditions including respiratory and cardiovascular diseases [2], while also exerting broader impacts on ecosystems, such as contributing to climate change and accelerating biodiversity loss. Against this backdrop, the accurate characterization, understanding, and modelling of air pollution have garnered increasing attention in the scientific community.

To comprehensively understand air quality, China has built a modern environmental monitoring network covering major cities and key regions for real-time dynamic observation of air pollutants. However, collected datasets often have missing values due to instrument malfunction, extreme weather, data transmission interruptions, and human errors [3]. Notably, data come from geographically distributed monitoring stations with substantial spatial heterogeneity, resulting in temporal and spatial missing data that impair analysis-reliable data integrity and continuity. Missing data severely limits accurate regional pollution assessment [4], hinders source attribution, early warning, forecasting, and environmental policy evaluation. Particularly for high-precision air quality prediction models, incomplete data causes insufficient or biased training, undermining output reliability [5]. Thus, effective data imputation is indispensable to reconstruct continuously missing spatio-temporal pollutant data.

Missing data in environmental data is generally unavoidable, and correct handling and filling of these gaps can enhance downstream research. Currently, such methods are divided into traditional methods and advanced methods based on deep learning. Traditional missing value processing methods can be broadly categorized into two main types: data deletion methods and statistical feature-based imputation techniques. However, these approaches exhibit substantial limitations when applied to air quality data, which often possess complex spatio-temporal characteristics. Although data deletion is straightforward to implement, it disrupts the continuity of time series, potentially leading the model to erroneously interpret discontinuous observations as continuous sequences, thereby introducing systematic bias [6]. Statistical imputation methods (e.g., mean/median imputation, linear imputation) consider only the statistical properties of individual variables, while neglecting key temporal dependencies such as trends and seasonality, as well as inter-variable relationships [7]. Moreover, these methods fail to capture spatial correlations, despite the fact that significant spatial autocorrelation typically exists between neighboring monitoring stations in air quality datasets.

In recent years, deep learning has shown great promise in missing value interpolation. Methods based on RNNs [8], CNNs [9], Transformers [10], and GANs [11] have achieved notable success in temporal data imputation. Transformers have become a focus due to their self-attention mechanism for long-range dependencies and inter-variable interactions, but they do not always outperform traditional models [12], as their permutation-invariant structure may lose temporal order information. To address this, improvements like RevIN [13] enhance adaptability via adaptive normalization. Meanwhile, the emerging KAN [14], based on function approximation theory, effectively captures temporal dynamics. The complementarity of Transformers (long-range dependence) and KAN (high-precision function learning) enables a more expressive spatio-temporal interpolation framework. Based on this, this paper proposes the KT model (KAN-Transformer) for air quality missing value imputation, integrating KAN and Transformer to reconstruct data from temporal dependence, inter-pollutant correlation, and sensor spatial distribution. Targeting air quality data's unique characteristics (missing values, spatio-temporal structure, nonlinearity, high-dimensional dependencies), a targeted framework is designed: 1) Transformer for global variable dependencies and temporal feature extraction; 2) MoE mechanism integrated into KAN to handle nonlinear, multiscale pollution evolution; 3) CSPM to dynamically model variables independently or jointly based on spatial heterogeneity; 4) RevIN to improve generalization by stabilizing feature distributions. These components form a cohesive architecture for complex air quality imputation. The main contributions of this paper are as follows:

1) This is the first study, to the best of our knowledge, to integrate the Kolmogorov-Arnold Network (KAN) with a Transformer-based architecture for missing value imputation in air quality data. The proposed model jointly captures temporal, inter-variable, and spatial dependencies, and demonstrates competitive performance.

2) A hybrid expert mechanism, combined with Reversible Instance Normalization (RevIN), is introduced to guide the model in learning nonlinear transformation patterns among pollutants, while enhancing its adaptability to distributional shifts in real-world datasets.

3) Using a channel selection mechanism based on Spearman's correlation coefficient, the channel input method is adaptively selected according to spatial correlation.

2. Related Work

In recent years, various imputation methods have been applied to missing values in air quality data, with the choice of methods often having a direct impact on the validity of subsequent analyses.

Traditional imputation approaches are primarily based on statistical principles, estimating missing values by analyzing the distributional characteristics of the data or the relationships among variables. Common techniques include mean im-

putation, median imputation, and mode imputation. These methods are easy to implement but typically focus solely on global statistical properties, thereby neglecting temporal dependencies among data points and latent correlation structures between variables within the time series.

With the advancement of deep learning technologies, significant breakthroughs have been achieved in the task of missing value imputation. Compared with traditional approaches, deep learning models exhibit superior nonlinear modeling capabilities and representation learning capacities, enabling them to better accommodate complex data structures and distributional shifts [15] [16]. Representative frameworks include models based on Recurrent Neural Networks (abbr. RNNs) [17], Convolutional Neural Networks (abbr. CNNs) [18], Graph Neural Networks (abbr. GNNs) [19] [20], and attention mechanisms [21]. In the context of time series imputation, the SAITS model [20] simulates real-world missing scenarios through a diagonal masking strategy and enhances imputation performance by integrating it with a reconstruction-based learning framework. ImputeFormer [22] combines a low-rank assumption with the self-attention mechanism of the Transformer architecture to efficiently impute high-dimensional spatiotemporal data. Meanwhile, multimodel fusion strategies have gained increasing attention due to their ability to integrate the strengths of different architectures, enabling more comprehensive extraction of the structural features in the data. For instance, TimesNet [23] introduces the chunked time-domain mixing (abbr. TSMix) mechanism, which fuses local convolution with global modeling capabilities to effectively capture multiscale temporal dependencies. ST-SILM [24] leverages a combination of Single Exponential Smoothing (abbr. SES) and Inverse Distance Weighting (abbr. IDW) with Long Short-Term Memory (abbr. LSTM) networks to achieve spatiotemporal correlation fusion. DSSE further improves imputation accuracy and robustness by integrating the global context modeling ability of the Transformer with the temporal feature extraction capability of Bidirectional LSTM (abbr. Bi-LSTM).

It is noteworthy that the Kolmogorov–Arnold Network (KAN), as a novel neural network architecture, has gradually demonstrated distinct advantages in time series modeling due to its highly flexible function approximation capabilities. For instance, TimeKAN [25] addresses the problem of multifrequency mixing by decomposing sequences into different frequency components, enabling more effective temporal modeling. WormKAN [26] introduces a concept-aware module that can efficiently detect and learn complex patterns such as concept drift. In addition, several studies have explored hybrid architectures that integrate Convolutional Neural Networks, Bidirectional Long Short-Term Memory networks (abbr. Bi-LSTMs), and KANs for tasks such as greenhouse gas emission prediction, demonstrating strong performance in capturing nonlinear dynamic changes [27]. Although KAN has shown promising results across various time series prediction tasks, its potential in missing value imputation has not yet been thoroughly explored yet. Therefore, investigating the applicability of KAN in this domain—par-

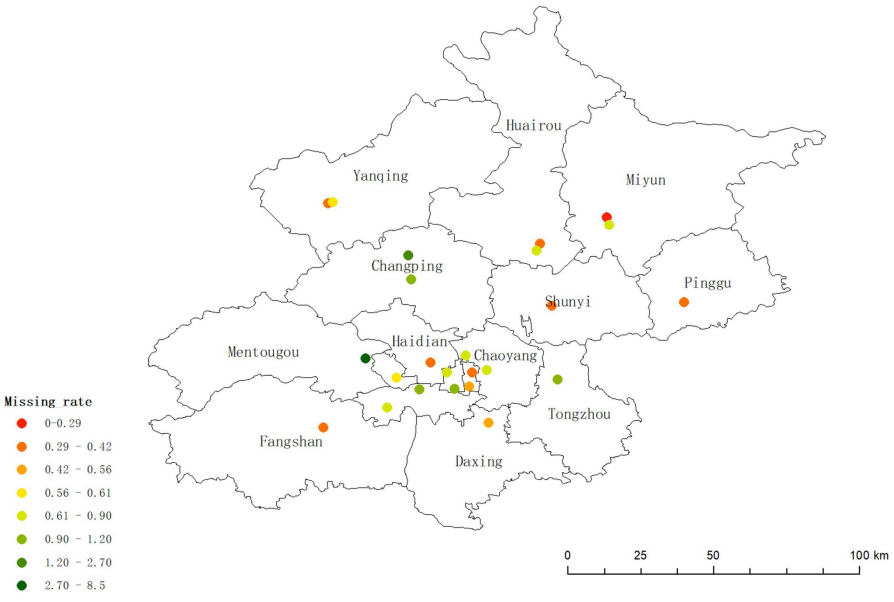
ticularly through its integration with attention mechanisms and hybrid expert architectures—has significant theoretical and practical value.

3. Materials and Methods

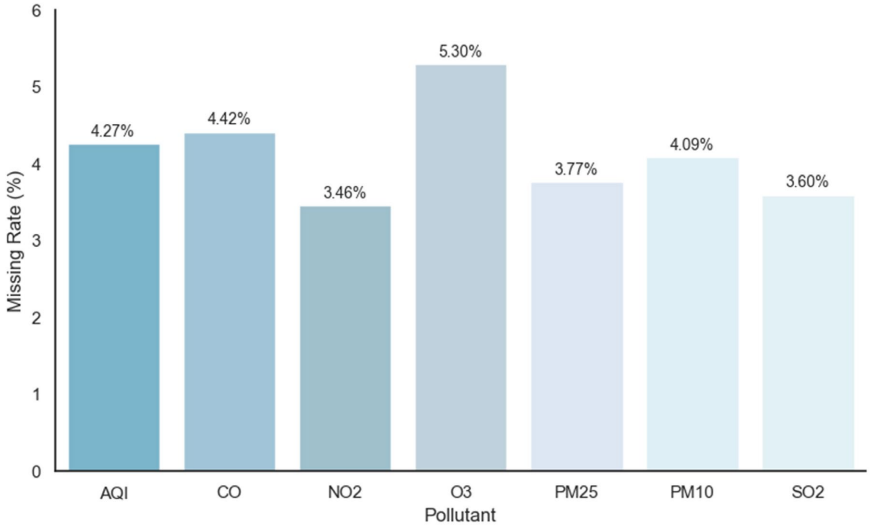
In this section, we first present a real-world air quality monitoring dataset. We then formulate the fundamental concepts and mathematical representation of the air quality data interpolation task by analyzing the inherent characteristics of the dataset. Finally, we present the overall architecture and functions of the proposed model in Section 3.5, and provide a detailed description of each module with the corresponding mathematical derivations.

3.1. Datasets

Experimental data were from the National Urban Air Quality Real-Time Release Platform (https://air.cnemc.cn:18007/), covering Jan 1, 2023, to Oct 12, 2024, including key pollutants ($PM_{2.5}$, PM_{10}, SO_2) as a multivariate time series with temporal dynamics, inter-variable correlations, and spatial heterogeneity (see **Figure 1(a)**). Analysis of Beijing's air quality data shows varying missing data across stations and pollutants (see **Figure 1(a)**, **Figure 1(b)**). "Missing amount" refers to unavailable hours, "missing rate" to the proportion of missing hours. Most stations have an annual missing rate of 0.4% - 1.2% (up to 8.5% for individual stations), with an average of over 3% across pollutants. The dataset is split: test set (Mar, Jun, Sep, Dec), validation set (Apr, Jul), training set (remaining months). To evaluate the model, artificial missing masks are applied to observed data to simulate different missing patterns and rates, ensuring reliable assessment of imputation accuracy and robustness.



(a) Spatial distribution and missing rate of monitoring stations.



(b) Proportion of missing air quality monitoring indicators.

Figure 1. Missing data by site and by pollutant. (a) Spatial distribution and missing rate of monitoring stations: with green to red indicating a gradual increase in the rate of missing data (This map is produced based on the standard map with review number GS(2019)3333 downloaded from the Standard Map Service website of the Ministry of Natural Resources, and the map has not been modified). (b) Proportion of missing.

3.2. Problem Formulation

Based on the characteristics of air quality monitoring data, we provide the following formal definition. Given a multivariate time series $X \in R^{T\times D}$, where $X = \{x_1, x_2, \cdots, x_T\}$ represents observations at T time steps. Each observation $x_t \in R^D$ at time step t contains D features corresponding to the pollutant concentrations measured at various stations, denoted as $\{x_t^1, \cdots, x_t^d, \cdots, x_t^D\} \in R^{1\times D}$. Due to circumstances beyond our control in the real-world, the dataset contains missing values, and the objective is to accurately estimate these missing entries. To represent the pattern of missingness in X, we introduce a binary mask matrix $M = \{m_1^d, m_2^d, \cdots, m_T^d\}$, $M \in R^{T\times N}$, define m_t^d and each $m_t \in \{0,1\}^N$ denotes the mask vector at time step t. The entries in M are defined as:

$$m_t^d = \begin{cases} 1 & \text{if } x_t^d \text{ is observed} \\ 0 & \text{if } x_t^d \text{ is missing} \end{cases} \tag{1}$$

This mask is used to guide the model in distinguishing between observed and unobserved entries during the imputation process. Where the value of m_t^d is 0 if the value of the dth variable x_t^d is missing, and the value of m_t^d is 1 if x_t^d is observable.

3.3. Masking Strategy

Since the true values of the original missing data are unavailable, evaluating model performance requires an artificial masking scheme during training to simulate realistic missing patterns as closely as possible. We assume that the data distribution

before and after missingness remains approximately consistent, allowing the estimation accuracy on artificially masked entries to serve as a proxy for evaluating the model's ability to recover the original missing values. The formula is expressed as follows:

$$\mathcal{L} = \frac{1}{\sum A_t^d} \sum_{t,d} A_t^d \cdot \ell\left(X_t^d, \hat{X}_t^d\right) \tag{2}$$

In this equation, X_t^d represents the original observation data, $\hat{X}_t^d$ represents the estimated values of the model, A_t^d represents the artificial masking matrix, and $\ell(\cdot)$ represents a defined loss function.

To differentiate artificial missingness from natural missingness, we define the artificial masking task as follows:

$$A_t^d = \begin{cases} 1 & \text{if } X_t^d \text{ is artificially masked} \\ 0 & \text{otherwise} \end{cases} \tag{3}$$

Let $A_t^d \in \{0,1\}^{T \times D}$ denote a binary mask matrix, where $A_t^d = 1$ marks artificially masked entries and $A_t^d = 0$ indicates retained data.

3.4. Methodology

The Kolmogorov-Arnold Network (abbr. KAN) is a novel neural network architecture inspired by the Kolmogorov-Arnold representation theorem, which demonstrates that any continuous multivariate function can be decomposed into a finite composition of univariate functions. This theoretical foundation is expressed mathematically as follows:

$$f\left(x_1, x_2, \cdots, x_n\right) = \sum_{q=1}^{2n+1} \Phi_q \left(\sum_{p=1}^{n} \varphi_{p,q}\left(x_p\right) \right) \tag{4}$$

where both Φ_q and $\varphi_{q,p}$ are nonlinear functions.

Based on this theorem, complex dependencies in high-dimensional input data can be effectively represented by compositions of nonlinear univariate functions. Recent advances in lightweight time series architectures have shown a clear shift from linear models to MLP-based designs, highlighting the effectiveness of nonlinear modeling. While MLPs rely on fixed activation functions at the nodes, Kolmogorov-Arnold Networks (abbr. KANs) introduce learnable activations along the edges, which grants them richer nonlinear expressivity. This edge-based flexibility makes KANs particularly well-suited for capturing intricate dependencies in high-dimensional data. It positions them as a compelling alternative to MLPs in a wide range of applications. The composite mapping of a KAN with L layers can be formally expressed as follows:

$$\mathrm{KAN}(x) = \left(\Phi_{L-1} \odot \Phi_{L-2} \odot \cdots \odot \Phi_0\right)(x) \tag{5}$$

where Φ_L corresponds to the function matrix of layer L and each matrix element is a trainable univariate function. The operator $\odot$ denotes function composition across layers, applied in order from Φ_0 to Φ_{L-1}.

KAN has demonstrated significant advantages in fields such as time series prediction and pattern recognition. In this study, it is first applied to the task of imputing missing values in spatiotemporal data, mainly based on the following considerations: 1) air quality data has complex spatiotemporal correlation, which requires strong nonlinear modeling capability; 2) the missing patterns of monitoring data often show non-uniform distribution, which requires the model to have the ability to capture local features.

3.5. Model Architecture

In response to the multidimensional spatiotemporal characteristics of air quality data, this study proposes a KAN-Transformer hybrid architecture, in short for KT designed to achieve high-precision imputation of missing values. KT simultaneously captures temporal dependencies, inter-variable correlations, and spatial heterogeneity by integrating three complementary modules: a hybrid expert KAN layer, a variable-wise Transformer module, and a channel-wise attention selection mechanism (ASPM). As illustrated in Figure 2, these components collectively form an end-to-end framework for spatio-temporal feature representation and reconstruction. To further enhance model robustness in non-stationary environments, a Reversible Instance Normalization (abbr. RevIN) strategy is applied prior to model training. This preprocessing step reduces the adverse effects of distributional shifts across variables and time periods, thereby enabling the model to operate on a more stable and homogeneous input space.

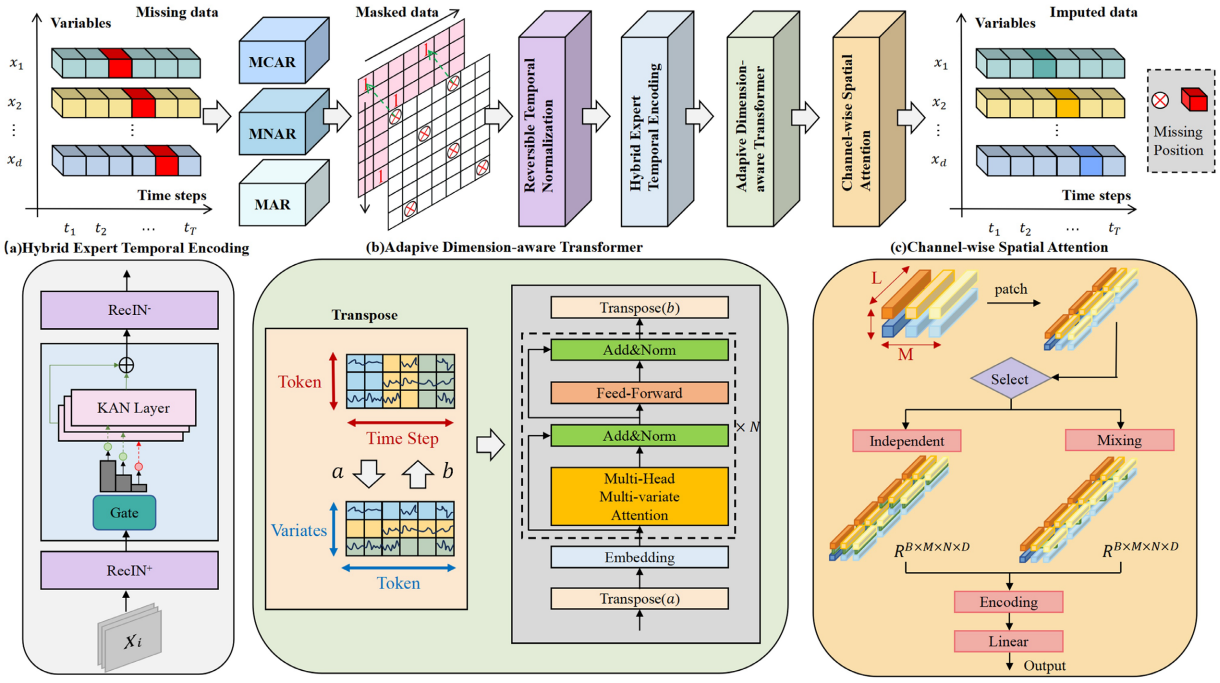


Figure 2. The overall architecture of our proposed model, as well as the (a) Hybrid expert temporal encoding, (b) Adaptive Dimension-aware Transformer, and (c) Channel-wise spatial attention.

3.5.1. Reversible Temporal Normalization

To address the pervasive non-stationarity in air quality data, this study introduces a Reversible Instance Normalization (abbr. RevIN) strategy prior to model training. This mechanism initially transforms the input sequences into a stationary distribution via instance normalization, effectively mitigating the impact of distributional shifts on model learning and enabling the network to better capture intrinsic temporal dependencies. During the output phase, the nonstationary components removed during normalization are seamlessly reintegrated through an inverse transformation, ensuring that the final predictions remain consistent with the original data distribution. As illustrated in Figure 3, this approach not only preserves the statistical characteristics of the raw data but also enhances the model's capacity to learn meaningful temporal patterns.

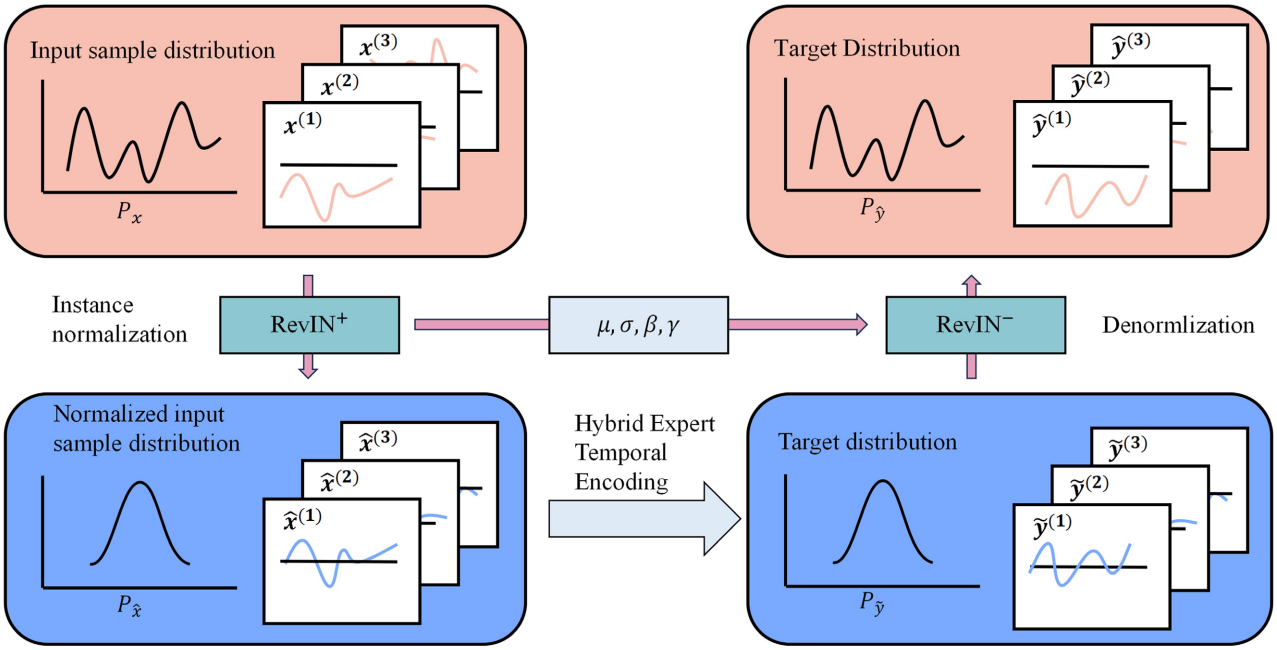


Figure 3. RevIN structure.

We combine the original data X with the generated mask matrix M to obtain x_M as an input and first calculate the mean average and variance of the inputs and then normalise them:

$$E_t\left[x_M\right]=\frac{1}{T}\sum_{j=1}^{T}x_{Mj} \tag{6}$$

$$Var\left[x_M\right]=\frac{1}{T}\sum_{j=1}^{T}\left(x_{Mj}-E_t\left[x_M\right]\right)^2 \tag{7}$$

$$\hat{x}_M=\gamma\left(\frac{x_M-E_t\left[x_M\right]}{\sqrt{Var\left[x_M\right]+\varepsilon}}\right)+\beta \tag{8}$$

where γ and β are both learnable parameters, ε is a constant, and T is

the time step length.

Subsequently, the normalised data $\hat{x}_M$ is fed into the hybrid expert KAN model and the model output is inversely normalised mapped to the original distribution space to obtain the output $\hat{y}_M$. This is done as follows:

$$\hat{y}_M = \sqrt{Var[x_M] + \varepsilon} \cdot \left(\frac{\tilde{y} - \beta}{\gamma} \right) + E_t[x_M] \tag{9}$$

Finally, the result of inverse normalization is linearly mapped back to the original feature dimensions and the missing is replaced with the resulting estimate to obtain the output $\hat{X}_1$ of the first module:

$$X_1 = W_1 \hat{y}_M + b_1 \tag{10}$$

$$\hat{X}_1 = M \odot X_1 + (1 - M) \odot X_1 \tag{11}$$

3.5.2. Hybrid Expert Temporal Encoding

As the initial processing layer of the model, the hybrid expert KAN layer effectively captures the complex distributions of pollutant concentrations through a collaborative multi-expert mechanism. This module, built upon the Kolmogoro v-Arnold Network hybrid expert framework, is specifically tailored to extract multilevel temporal features from preprocessed input data. Unlike traditional multilayer perceptrons (MLPs) that rely on fixed functional mappings, the KAN network leverages a combination of learnable univariate basis functions, enabling superior modeling of intricate nonlinear relationships. Inspired by [27], this study designs a cooperative learning architecture comprising four distinct types of expert functions (see **Table 1**) to accommodate the multiscale dynamics and heterogeneity inherent in air quality monitoring data.

The model consists of a KAN network with a hybrid layer of experts, which internally contains a gated network that assigns the KAN layer to the corresponding variables based on the read time series features, and the KAN internally contains different spline functions, *i.e.*, each expert is responsible for processing the input data differently. Therefore, the KAN hybrid expert network with N experts can be represented as:

$$x' = \sum_{i=1}^{N} g(\hat{x}_M)_i \, K_i(\hat{x}_M) \tag{12}$$

where $g(\cdot)$ is the gating network. $K(\cdot)$ is the KAN. x' is the output. This hybrid expert structure can effectively address the heterogeneity of data distribution and learn features from different variables for better missing value estimation.

Each expert applies a different function for fitting, respectively.

The gating network is a critical component of the architecture, as it calculates the weights for each expert based on the input data. The *Softmax* function is then applied to normalize these weights into a probability distribution. This enables the dynamic selection of the most suitable expert network for computation, based on the varying characteristics of the data.

Table 1. Formulas for different fitting functions and their meanings.

Expert Name	Function Form	Application Scenario
Mexican Hat Wavelet Expert	$\psi = \left(1-t^2\right)\mathrm{e}^{-\frac{1}{2}t}$	It is sensitive to sudden pulses, amplifies local signal features, and provides a good fit to sudden increases and transient spikes in pollutant levels.
Morlet Wavelet Expert	$\psi(t) = \pi^{-\frac{1}{4}}\mathrm{e}^{i w_0 t}\mathrm{e}^{-\frac{t^2}{2}}$	Suitable for analyzing the periodic variations in air pollutant concentrations.
Derivative of Gaussian (DOG) Wavelet Expert	$\psi(t) = -\frac{t}{\sigma^2}\mathrm{e}^{-\frac{t^2}{2\sigma^2}}$	Rapid increase or decrease in concentration of applicable pollutant.
B-spline	$\psi(t) = \beta_0 + \sum_{p=1}^{p}\sum_{i=0}^{n_p}\beta_{p,i}N_{p,i,k_p}\left(t_p\right) + \varepsilon$	Rapid increase or decrease in concentration of applicable pollutant.

$$g_{Softmax}(x) = Softmax\left(xw_g\right) \tag{13}$$

To learn the relevant features of the task more efficiently, the network adds Gaussian noise to the input time series during training and selects only the top k best matching experts for computation.

$$g_{Softmax}(x) = Softmax\left(KeepTopK\left(H(x)\right), k\right) \tag{14}$$

$$H(x) = xw_g + Norm\left(Softplus\left(xw_{noise}\right)\right) \tag{15}$$

where $Norm(\cdot)$ denotes standard normalization and w_{noise} is Gaussian noise. $KeepTopK(\cdot, k)$ retains the top k largest elements of the input vector $H(x)$. This ensures that the Softmax operation is computed only over the top k values, effectively enforcing sparsity and enabling selective attention.

The architecture achieves adaptive synergy among experts through a dynamic gating mechanism, which maintains the specialized feature extraction capability of each expert function and improves the generalization performance of the model through integrated learning. This design provides a new solution for the time-series feature analysis of complex pollution processes.

3.5.3. Adaptive Dimension-Aware Transformer

Traditional Transformer-based multivariate time series methods usually embed all variables' data at one time step into a single representation [28], which captures temporal dependencies but assumes tight coupling of all variables. However, air quality datasets often have temporal misalignment among pollutants due to lagged pollution event responses, leading to mutual interference and degraded feature correlation capture when embedding multiple variables into one representation. Recent studies [29] show that input dimension transposition (treating each variable as an independent token and modeling the feature dimension explicitly) can mitigate this issue. Inspired by this, we propose a feature-oriented transposed Transformer, shifting from time-domain to feature-centric tokenization. Specifically, the input tensor is transposed to form an independent token sequence for

each pollutant, projected to hidden dimensions via linear mapping to get refined representation α, and then attention mechanisms and feed-forward networks are applied along the feature dimension to capture inter-variable correlations while preserving temporal structure. This enhances global multivariate pattern extraction, facilitating more accurate missing value imputation.

$$X^{\mathrm{T}} = Invert\left(\hat{X}_1\right) \tag{16}$$

$$M^{\mathrm{T}} = Invert\left(M\right) \tag{17}$$

$$\alpha = W_2 Concat\left(X^{\mathrm{T}}, M^{\mathrm{T}}\right) + b_2 \tag{18}$$

where W_2, b_2 is a learnable parameter.

Secondly, the multi-head self-attention of the mask is utilized with the N stacked layers of the feed-forward neural network to operate self-attention on the inputs to effectively capture the feature correlations, thus improving the representation and learning performance. To further learn the feature representation, we apply two linear projections and add ReLU activation between them to obtain γ, as deeper structures can generally learn better.

$$\beta = FeedForward\left(MHA\left(\alpha\right)\right)^N \tag{19}$$

$$\gamma = W_4 ReLU\left(W_3\beta + b_3\right) + b_4 \tag{20}$$

where *MHA* denotes the multi-head self-attention mechanism. Finally, the learned feature association information is reintegrated with the original temporal structure through an inversion operation, the output of which is:

$$X_2 = Invert\left(\gamma\right) \tag{21}$$

$$\hat{X}_2 = M \odot X + \left(1 - M\right) \odot X_2 \tag{22}$$

3.5.4. Channel-Wise Spatial Attention

The previous inverse Transformer adopted a fully channel-independent approach (processing each variable separately). While it effectively eliminates inter-variable interference during normalization and reduces noise, it sacrifices inter-variable dependency modeling, which is critical for capturing complex pollutant interactions. In air quality data, variable interdependencies reflect important spatial relationships and co-evolving pollution dynamics across monitoring sites; ignoring them limits the model's ability to utilize data structure. Thus, we propose a selective channel-based spatial attention mechanism for spatio-temporal feature extraction. Unlike the inverse Transformer's pure channel independence, it adaptively chooses channel-independent or mixing strategies by assessing inter-variable correlation via Spearman correlation coefficient. Its motivation is twofold: 1) Preserve noise robustness (maintain channel independence for weak correlations to avoid noise and overfitting); 2) Enhance information sharing (adopt channel mixing for strong correlations to improve spatial dependency modeling).

Patching Local embedding starts with transforming input $\hat{x}_2' \in R^{B\times L\times M}$ to $Z \in R^{B\times M\times L}$, then slicing in the time dimension to get $z_p \in R^{B\times M\times N\times P}$ (P = window size, N = number of windows). Each variable sequence $x_{1:L}^i$ is split into $p^i \in R^{N\times P}$; with step S, $N = \left[\frac{L-P}{S}+1\right]$. The split data is mapped to hidden dimensions to obtain $Z_{emb} = Emb\left(Z_p\right) \in R^{B\times M\times N\times D}$.

Channel Selection Channel selection is performed using Spearman correlation coefficient (a non-parametric metric for monotonic relationships, calculated via rank to reduce outlier impact). We first extract training dataset $X = \{x_1, x_2, \cdots, x_N\}$, compute Spearman coefficients p_{ij} between x_i and x_j ($i \neq j$), then count correlated sequences under different thresholds: K_i^λ (high correlation) and K_i^0 (non-negative correlation).

$$K_i^\lambda = \sum_{j=1}^{M} 1\left(\rho_{i,j} \geq \lambda\right) \tag{23}$$

$$K_i^0 = \sum_{j=1}^{M} 1\left(\rho_{i,j} \geq 0\right) \tag{24}$$

$$r = \frac{\rho_{\max}^\lambda}{\rho_{\max}^0} \tag{25}$$

where λ is a human-set threshold. $1(\cdot)$ is an indicator function that takes the value of 1 when the condition is met and 0 otherwise. Then, the number of high correlations and the number of non-negative correlations are $\rho_{\max}^\lambda = \max\left(K^\lambda\right)$ and $\rho_{\max}^0 = \max\left(K^0\right)$, respectively. And calculate the high correlation ratio, defined as $r = \frac{\rho_{\max}^\lambda}{\rho_{\max}^0}$. Selection of channel selection strategies based on high correlation ratios:

$$C = \begin{cases} 1 & \text{if } r \geq 1-\lambda \\ 0 & \text{otherwise} \end{cases} \tag{26}$$

where if $C = 1$ we adopt a channel mixing strategy and vice versa a channel independent strategy.

The embedding of the data in front of the input encoder is different for different channel choices. If channel independence is chosen, the different channels are independent of each other, *i.e.*, the model processes each channel separately without sharing or exchanging information, and the data is reshaped as: $X_{ind} \in R^{(B\times M)\times N\times D}$, which will be obtained by encoding them separately in the encoder:

$$H_{ind} = Encoder\left(X_{ind}\right), H_{ind} \in R^{(B\times N)\times M\times D} \tag{27}$$

Instead the channel data can interact with each other by reshaping the data as: $X_{mix} \in R^{(B\times N)\times M\times D)}$, the variable dimensions are preserved, at which point multiple channels are fed together into the encoder, at which point the self-attention in the Encoder allows the information between different channels to interact with each other:

$$
\begin{aligned}
&H_{mix} = Encoder\left(X_{mix}\right), H_{mix} \in R^{(B\times N)\times M\times D} \\
&H_{out} = rearrange\left(H_{ind}, (BM)ND \to BMND\right) \\
&H_{out} = rearrange\left(H_{mix}, (BN)MD \to BMND\right)
\end{aligned} \tag{28}
$$

In order to map the results of the above spatial dimensionality processing to the target interpolation format, we exchange the dimensions of the encoder outputs to obtain $H \in R^{B\times N\times M\times D}$. If the channels are independent, each variable channel is interpolated individually, and different variables are mapped using separate linear layers with linear transformations, and different channels do not share weights. Finally, we stack all channels; the process is as follows:

$$Z_c = Flatten(H) = reshape(H) \in R^{B\times(D\times N)} \tag{29}$$

$$y_c = W_c z_c + b_c, y_c \in R^{B\times H} \tag{30}$$

$$y = Stack\left(\left[y_1, y_2, \cdots, y_C\right], dim = 1\right), y \in R^{B\times C\times H} \tag{31}$$

where $W_c \in R^{H\times(D\times N)}$ is a channel c independent weight matrix and $b_c \in R^H$ is a bias term.

If the channels are mixed, all channels go through the same linear layer together and all parameters are shared, denoted as follows:

$$Z = Flatten(H) = reshape(H) \in R^{B\times C\times(D\times N)} \tag{32}$$

$$y = WZ + b, y \in R^{B\times C\times H} \tag{33}$$

where $W \in R^{H\times(D\times N)}, b \in R^H$.

Finally, the final imputation results are generated:

$$\hat{X}_3 = M \odot X + (1 - M) \odot y \tag{34}$$

4. Experiments

4.1. Experimental Setup

All experiments are conducted on workstations equipped with NVIDIA RTX 3060 GPUs. The model is trained using the Adam optimizer implemented in PyTorch. The learning rate was set to 0.001, and the hyperparameters are detailed in **Table 2**. To evaluate the model's performance, three commonly used metrics are employed to quantify the performance: mean absolute error (MAE), mean relative error (MRE), and root mean square error (RMSE), with their respective formulas defined as follows:

$$\mathrm{MAE}\left(\hat{X}, X, M\right) = \frac{\sum_{d=1}^{D}\sum_{t=1}^{T}\left|\left(\hat{X} - X\right)\odot M\right|_t^d}{\sum_{d=1}^{D}\sum_{t=1}^{T} M_t^d} \tag{35}$$

$$\mathrm{RMSE}\left(\hat{X}, X, M\right) = \sqrt{\frac{\sum_{d=1}^{D}\sum_{t=1}^{T}\left(\left(\left(\hat{X} - X\right)\odot M\right)^2\right)_t^d}{\sum_{d=1}^{D}\sum_{t=1}^{T} M_t^d}} \tag{36}$$

$$\mathrm{MRE}\left(\hat{X}, X, M\right) = \frac{\sum_{d=1}^{D}\sum_{t=1}^{T}\left|\left(\hat{X} - X\right)\odot M\right|_t^d}{\sum_{d=1}^{D}\sum_{t=1}^{T}\left|\hat{X}\odot M\right|_t^d} \tag{37}$$

where $\hat{X}$ denotes the estimated value of the air quality data, X denotes the initial value of the air quality data, M denotes the mask matrix, D denotes the variable, and T denotes the time.

Table 2. Hyper-parameters for model training.

Hyper-params	Air-Quality
Batch_size	64
d_model	256
n_head	4
n_group	2

4.2. Baseline Methodology

To evaluate the performance of the proposed model, we compare it against several deep learning-based models (*i.e.*, SAITS, Transformer, TimesNet, GP-VAE, BRITS, MRNN, ImputeFormer, ITransformer, TEFN, TimeMixer) and statistical methods (*i.e.*, Mean and Median) with similar methodological backgrounds. All methods used the same masking strategy, data partitioning approach, and evaluation protocol. 1) Mean 2) Median 3) SAITS [30]: A diagonal mask-based bi-level self-attention framework for missing value imputation. 4) Transformer [31]: Imputes missing values using the encoder module of the Transformer architecture. 5) TimesNet [23]: A method combining CNN and self-attention mechanisms that transforms time series into 2D tensors and applies 2D convolutional kernels for feature extraction. 6) GP-VAE [32]: Integrates the principles of variational autoencoders and Gaussian processes by embedding multivariate time series with missing values into a low-dimensional latent space, where Gaussian processes model the temporal dynamics. 7) BRITS [33]: Employs bidirectional recurrent neural networks (RNNs), leveraging both historical information and feature regression for missing value estimation. 8) MRNN [34]: Utilizes multidirectional RNNs to estimate missing values in time series data. 9) ImputeFormer [22]: A spatio-temporal imputation model that combines low-rank induction with Transformer-based attention mechanisms. 10) ITransformer [29]: Performs data imputation by learning feature dependencies through a feature-wise attention mechanism applied to transposed time series data. 11) TEFN [35]: Imputation missing values using a lightweight neural network model based on evidence theory and information fusion. 12) TimeMixer [36]: Captures complex temporal patterns for interpolation through a multi-scale, multi-resolution attention mechanism.

4.3. Imputation Performance

To assess the generalizability of the proposed model across diverse scenarios, we

compare its performance against various baseline methods under different missing rates and missing patterns. The imputation results, evaluated using mean absolute error (MAE), mean relative error (MRE), and root mean square error (RMSE), are summarized as follows.

4.3.1. Imputation under MCAR

In real-world settings, sensor aging or failure often causes isolated missing values. We simulate such point-wise missingness with rates of 1%, 5%, and 10%. As shown in Table 3, statistical methods (Mean, Median) perform worst and degrade sharply as missingness increases, failing to capture air quality data dynamics. Deep learning methods are more stable. RNN-based models work reasonably at low missing rates but suffer from error accumulation under higher missingness. Among Transformer-based approaches, ImputeFormer and SAITS leverage self-attention for global temporal modeling; however, SAITS underperforms at 1% due to its focus on long sequences. TimesNet and TimeMixer show limited adaptability, while TEFN—designed for prediction—struggles with precise local reconstruction. GP-VAE exhibits robustness at high missing rates but still incurs notable errors. KT consistently outperforms all baselines, thanks to KAN's strong nonlinear modeling of complex pollutant patterns and its integration of global attention without error accumulation. This advantage is especially pronounced under high missingness. For a broader view, Figure 4 illustrates performance trends across missing rates from 1% to 70%.

Table 3. In the MCAR missing data scenario, the model imputation results show missing rates of 1%, 5%, and 10%. The best results are indicated in bold, and the second-best results are indicated with an underline.

Model/Missing Rate	1%	5%	10%
Metric	MAE/RMSE/MRE	MAE/RMSE/MRE	MAE/RMSE/MRE
Mean	0.8633/1.3401/1.0293	0.8723/1.4292/1.0277	0.8742/1.3941/1.0314
Median	0.8164/1.3824/0.9734	0.8231/1.4350/0.9711	0.8301/1.4704/0.9789
Transformer	0.2663/0.4321/0.2812	0.2763/0.4685/0.2901	0.2901/0.4871/0.3071
MRNN	0.3684/0.7114/0.4408	0.3868/0.7846/0.4546	0.3906/0.7677/0.4689
TimeMixer	0.3721/0.7901/0.5901	0.3918/0.8637/0.6117	0.4212/0.7306/0.5990
TEFN	0.3911/0.7308/0.4907	0.4075/0.7558/0.5000	0.4209/0.7708/0.5204
GP-VAE	0.3820/0.6142/0.4604	0.4151/0.7203/0.4970	0.4288/0.6798/0.5086
TimesNet	0.3414/0.6521/0.4115	0.3546/0.7555/0.4228	0.3580/0.7622/0.4201
SAITS	0.2272/0.4124/0.2763	0.2356/0.4584/0.2791	0.2353/<u>0.4630</u>/<u>0.2727</u>
Imputeformer	<u>0.1785</u>/0.4306/<u>0.2152</u>	0.2844/0.5500/0.3401	0.2877/0.5770/0.3459
Itransformer	0.2391/0.6236/0.2882	0.2573/0.6140/0.3077	0.2718/0.6597/0.3268
BRITS	0.1902/<u>0.3613</u>/0.2293	<u>0.2046</u>/<u>0.3849</u>/<u>0.2489</u>	<u>0.2279</u>/0.4897/0.2754
2*KT	**0.1446/0.2757/0.1758**	**0.1335/0.3026/0.1582**	**0.1378/0.3118/0.1657**
	↓19%/↓24%/↓18%	↓34%/↓21%/↓36%	↓39%/↓32%/↓39%

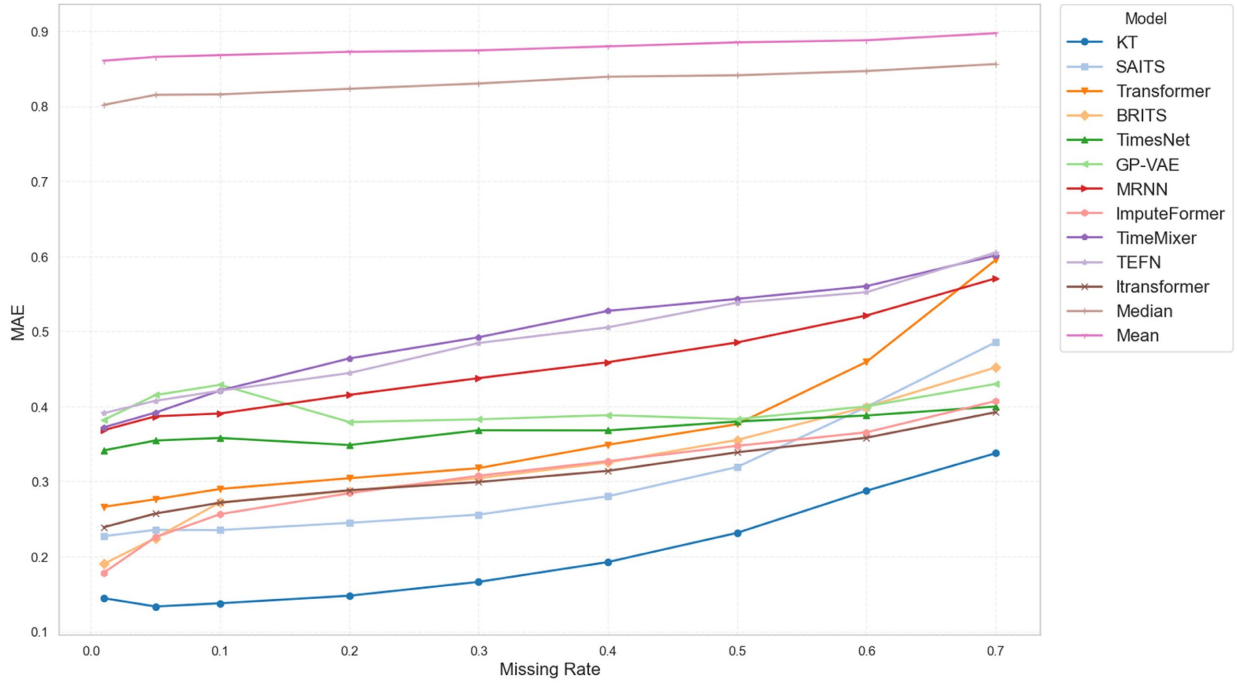
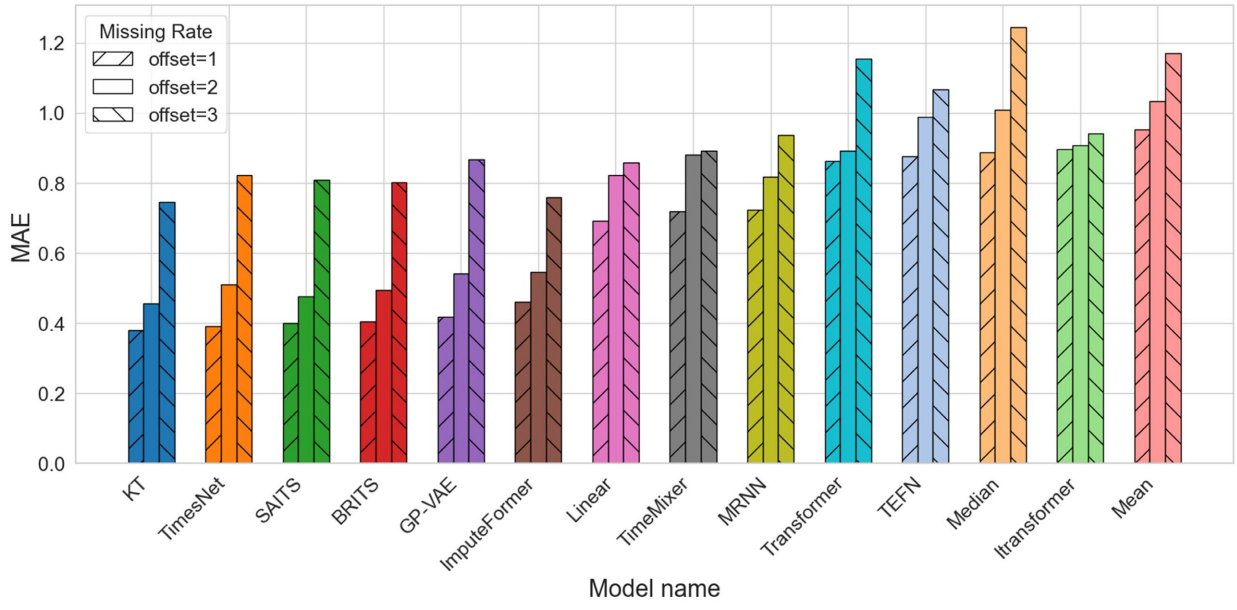


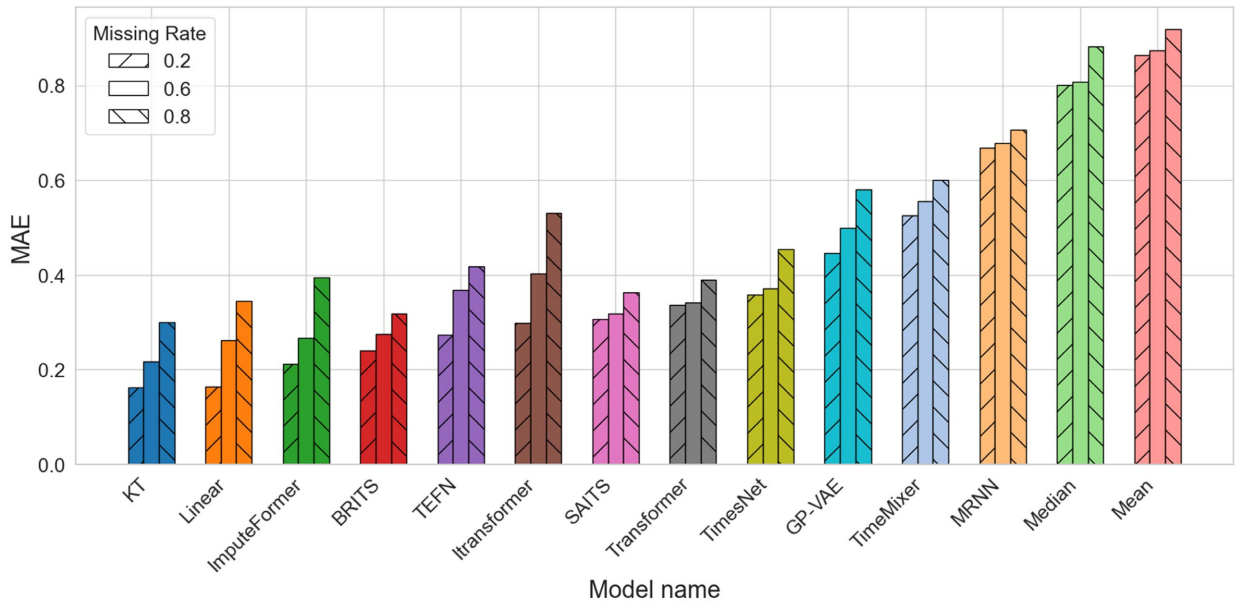
Figure 4. MCAR (missing rate from 1% to 70%).

4.3.2. Imputation under MNAR and MAR

Additionally, the results for simulated MNAR and MAR missingness are shown in Figure 5. Overall, deep learning methods consistently outperform statistical approaches across all missing rates. As the missing proportion increases, all methods exhibit varying degrees of performance degradation; however, deep learning models show a relatively milder decline, indicating stronger robustness to missing data. Among all methods, our model demonstrates the best overall performance.



(a) MNAR



(b) MAR

Figure 5. Performance of different models under MNAR and MAR missingness at various missing rates. (a) MNAR (b) MAR.

5. Ablation Experiment

In this section, we validate the effectiveness of KT architecture through four ablation experiments, each designed to validate the necessity of individual components. Specifically, we sequentially remove the KAN module, the transposed transformer module, and the channel-selection transformer module to assess their respective contributions. Furthermore, we verify the importance of our transposition operation by replacing the variable-dimension transformer with a time-dimension transformer, confirming that modeling along the variable dimension yields superior performance. As shown in Table 4.

1) Effectiveness of the KAN Architecture in Modeling Temporal Dependencies. We evaluate the capability of the KAN architecture in capturing temporal dependencies through comparative experiments. Specifically, we construct a unified input by applying a linear mapping to the transposed raw data and mask matrix, and feed it into the Transformer model operating along the variable dimension. The experimental results demonstrate that explicitly modeling temporal dependencies via the KAN architecture leads to a substantial improvement in interpolation performance, particularly under a 10% missing rate, where a 15.9% performance gain is observed. This enhancement may be attributed to the inherent periodic patterns in air quality data, which are effectively captured by the temporal modeling module embedded within the KAN architecture.

2) The Importance of Modeling Inter-Variable Dependencies. Table 4 presents a performance comparison after removing the variable-dimension Transformer module. The experimental results indicate that ignoring inter-variable interactions leads to a significant degradation in imputation performance, with RMSE increasing by 133%, even under conditions of a very low missing rate (1%). This

finding highlights the strong correlations among variables in air quality monitoring data (such as $PM_{2.5}$, SO_2, NO_2, etc.) and underscores the necessity of modeling cross-feature dependencies through our variable-dimension Transformer module.

3) Contribution of Spatial Information Modeling. In the ablation experiments targeting spatial information modeling, we remove the channel-selection self-attention mechanism and directly combined the outputs of the first two modules with the mask matrix to produce the final result. Although this simplified version exhibits relatively stable performance, its evaluation metrics are still notably inferior to those of the full model. This indicates that: 1) feature selection in the spatial dimension plays a critical role in imputation accuracy, and 2) relying solely on spatio-temporal features without explicitly modeling the spatial correlations within the sensor network leads to substantial information loss.

Table 4. Different versions of our method.

Model/Missing Rate	1%	5%	10%
NO_ONE	0.1327/0.2582/0.1615	0.1546/0.3231/0.1832	0.1640/0.3603/0.1971
NO_TWO	0.1468/0.3210/0.1764	0.1540/0.2901/0.1875	0.1640/0.3314/0.1944
NO_THREE	0.1448/0.2626/0.1761	0.1498/0.3101/0.1776	0.1546/0.3231/0.1832
KT	**0.1292/0.2934/0.1372**	**0.1335/0.3026/0.1582**	**0.1378/0.3118/0.1657**

6. Conclusion

In this study, an innovative imputation method for air quality data is proposed, which significantly improves data reconstruction performance under complex missing scenarios by systematically integrating information across three dimensions: temporal dependency, inter-variable correlation, and spatial correlation. The model employs a three-level modeling framework: first, the KAN hybrid expert architecture is used to capture the periodic characteristics and non-stationary trends of each pollutant; second, the nonlinear correlations among pollutants are modeled using the variable-dimension Transformer; and third, a spatial information fusion strategy is dynamically built based on the Spearman correlation coefficient. Experimental results show that the proposed method demonstrates strong adaptability and achieves performance improvements of 1.1% to 17.1% compared to the best baseline state-of-the-art (SOTA) model. Moreover, the model effectively controls computational complexity while maintaining high performance, with a 71.9% reduction in parameter count compared to SATIS, the strongest among the baseline models. Comparative experiments with 13 baseline methods confirm that our model achieves substantial improvements across three key metrics: imputation accuracy, stability, and generalization capability. The primary contribution of this study is the introduction of a three-dimensional coupled modeling paradigm that simultaneously captures temporal, feature-wise, and spatial dependencies, offering a novel methodological approach for complex spatio-temporal data restoration.

Data Availability

Data will be made available on request.

Conflicts of Interest

The author declares no conflicts of interest regarding the publication of this paper.

References

[1] Cohen, A.J., Brauer, M., Burnett, R., Anderson, H.R., Frostad, J., Estep, K., *et al.* (2017) Estimates and 25-Year Trends of the Global Burden of Disease Attributable to Ambient Air Pollution: An Analysis of Data from the Global Burden of Diseases Study 2015. *The Lancet*, **389**, 1907-1918. https://doi.org/10.1016/s0140-6736(17)30505-6

[2] Chen, Z., Liu, P., Xia, X., Wang, L. and Li, X. (2022) The Underlying Mechanism of $PM_{2.5}$-Induced Ischemic Stroke. *Environmental Pollution*, **310**, Article 119827. https://doi.org/10.1016/j.envpol.2022.119827

[3] Yu, Y., Yu, J.J.Q., Li, V.O.K. and Lam, J.C.K. (2017) Low-Rank Singular Value Thresholding for Recovering Missing Air Quality Data. 2017 *IEEE International Conference on Big Data* (*Big Data*), Boston, 11-14 December 2017, 508-513. https://doi.org/10.1109/bigdata.2017.8257965

[4] Chan, C.K. and Yao, X. (2008) Air Pollution in Mega Cities in China. *Atmospheric Environment*, **42**, 1-42. https://doi.org/10.1016/j.atmosenv.2007.09.003

[5] Ma, J., Cheng, J.C.P., Ding, Y., Lin, C., Jiang, F., Wang, M., *et al.* (2020) Transfer Learning for Long-Interval Consecutive Missing Values Imputation without External Features in Air Pollution Time Series. *Advanced Engineering Informatics*, **44**, Article 101092. https://doi.org/10.1016/j.aei.2020.101092

[6] Kaiser, J. (2014) Dealing with Missing Values in Data. *Journal of Systems Integration*, **5**, 42-51. https://doi.org/10.20470/jsi.v5i1.178

[7] Luo, Y., Cai, X., Zhang, Y., Xu, J., *et al.* (2018) Multivariate Time Series Imputation with Generative Adversarial Networks. *Proceedings of the* 32*nd International Conference on Neural Information Processing Systems*, 3-8 December 2018, Red Hook, 1603-1614.

[8] Bradbury, J., Merity, S., Xiong, C. and Socher, R. (2016) Quasi-Recurrent Neural Networks. arXiv:1611.01576.

[9] Long, J., Shelhamer, E. and Darrell, T. (2015) Fully Convolutional Networks for Semantic Segmentation. 2015 *IEEE Conference on Computer Vision and Pattern Recognition* (*CVPR*), Boston, 7-12 June 2015, 3431-3440. https://doi.org/10.1109/CVPR.2015.7298965

[10] Ma, J., Shou, Z., Zareian, A., Mansour, H., Vetro, A. and Chang, S.-F. (2019) CDSA: Cross-Dimensional Self-Attention for Multivariate, Geo-Tagged Time Series Imputation. arXiv:1905.09904.

[11] Goodfellow, I.J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A. and Bengio, Y. (2014) Generative Adversarial Nets. In: Ghahramani, Z., Welling, M., Cortes, C., Lawrence, N. and Weinberger, K., Eds., *NIPS*'14: *Proceedings of the* 28*th International Conference on Neural Information Processing Systems*, Vol. 2, MIT Press, 2672-2680.

[12] Zeng, A., Chen, M., Zhang, L. and Xu, Q. (2023) Are Transformers Effective for Time Series Forecasting? *Proceedings of the AAAI Conference on Artificial Intelligence*,

37, 11121-11128. https://doi.org/10.1609/aaai.v37i9.26317

[13] Kim, T., Kim, J., Tae, Y., Park, C., Choi, J.-H. and Choo, J. (2022) Reversible Instance Normalization for Accurate Time-Series Forecasting against Distribution Shift. *International Conference on Learning Representations*, Chicago, 25-29 October 2022. https://openreview.net/pdf?id=cGDAkQo1C0p

[14] Liu, Z., Wang, Y., Vaidya, S., Ruehle, F., Halverson, J., Soljačić, M., Hou, T.Y. and Tegmark, M. (2024) Kan: Kolmogorov-Arnold Networks. arXiv:2404.19756.

[15] Ravanelli, M., Brakel, P., Omologo, M. and Bengio, Y. (2018) Light Gated Recurrent Units for Speech Recognition. *IEEE Transactions on Emerging Topics in Computational Intelligence*, **2**, 92-102. https://doi.org/10.1109/tetci.2017.2762739

[16] Du, S., Li, T. and Horng, S. (2018) Time Series Forecasting Using Sequence-To-Sequence Deep Learning Framework. 2018 9*th International Symposium on Parallel Architectures, Algorithms and Programming* (*PAAP*), Taipei, 26-28 December 2018, 171-176. https://doi.org/10.1109/paap.2018.00037

[17] Nur'Adzan, N.A., Jaya, M.I., Faizal, M. and Zamri, N.A. (2025) A Systematic Review of Recurrent Neural Network Adoption in Missing Data Imputation. *International Journal of Computing and Digital Systems*, **17**, 1-17. https://doi.org/10.12785/ijcds/1571041166

[18] Yu, Y., Li, V.O.K., Lam, J.C.K., Chan, K. and Zhang, Q. (2025) CTDI: CNN-Transformer-Based Spatial-Temporal Missing Air Pollution Data Imputation. *IEEE Transactions on Big Data*, **11**, 2443-2456. https://doi.org/10.1109/tbdata.2025.3533882

[19] Cini, A., Marisca, I. and Alippi, C. (2021) Filling the Gaps: Multivariate Time Series Imputation by Graph Neural Networks. arXiv:2108.00298.

[20] Kim, S., Lee, T. and Lee, J. (2025) TMF-GNN: Temporal Matrix Factorization-Based Graph Neural Network for Multivariate Time Series Forecasting with Missing Values. *Expert Systems with Applications*, **275**, Article 127001. https://doi.org/10.1016/j.eswa.2025.127001

[21] Yldz, A.Y., Koç, E. and Koç, A. (2022) Multivariate Time Series Imputation with Transformers. *IEEE Signal Processing Letters*, **29**, 2517-2521. https://doi.org/10.1109/lsp.2022.3224880

[22] Nie, T., Qin, G., Ma, W., Mei, Y. and Sun, J. (2024) ImputeFormer: Low Rankness-Induced Transformers for Generalizable Spatiotemporal Imputation. *Proceedings of the* 30*th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, Barcelona, 25-29 August 2024, 2260-2271. https://doi.org/10.1145/3637528.3671751

[23] Wu, H., Hu, T., Liu, Y., Zhou, H., Wang, J. and Long, M. (2022) TimesNet: Temporal 2D-Variation Modeling for General Time Series Analysis. arXiv:2210.02186.

[24] Tan, S., Wang, Y., Yuan, Q., Zheng, L., Li, T., Shen, H., *et al.* (2022) Reconstructing Global $PM_{2.5}$ Monitoring Dataset from Openaq Using a Two-Step Spatio-Temporal Model Based on SES-IDW and LSTM. *Environmental Research Letters*, **17**, Article 034014. https://doi.org/10.1088/1748-9326/ac52c9

[25] Huang, S., Zhao, Z., Li, C. and Bai, L. (2025) Timekan: Kan-Based Frequency Decomposition Learning Architecture for Long-Term Time Series Forecasting. arXiv:2502.06910.

[26] Xu, K., Chen, L. and Wang, S. (2024) Are Kan Effective for Identifying and Tracking Concept Drift in Time Series? arXiv:2410.10041.

[27] Han, X., Zhang, X., Wu, Y., Zhang, Z. and Wu, Z. (2024) Kan4tsf: Are Kan and Kan-Based Models Effective for Time Series Forecasting? arXiv:2408.11306.

[28] Zhou, H., Zhang, S., Peng, J., Zhang, S., Li, J., Xiong, H., *et al.* (2021) Informer: Beyond Efficient Transformer for Long Sequence Time-Series Forecasting. *Proceedings*

of the AAAI Conference on Artificial Intelligence, **35**, 11106-11115. https://doi.org/10.1609/aaai.v35i12.17325

[29] Liu, Y., Hu, T., Zhang, H., Wu, H., Wang, S., Ma, L. and Long, M. (2024) iTransformer: Inverted Transformers Are Effective for Time Series Forecasting. arXiv:2310.06625. https://arxiv.org/abs/2310.06625

[30] Du, W., Côté, D. and Liu, Y. (2023) SAITS: Self-Attention-Based Imputation for Time Series. *Expert Systems with Applications*, **219**, Article 119619. https://doi.org/10.1016/j.eswa.2023.119619

[31] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł. and Polosukhin, I. (2017) Attention Is All You Need. *Advances in Neural Information Processing Systems*, **30**. https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[32] Fortuin, V., Baranchuk, D., Rätsch, G. and Mandt, S. (2020) GP-VAE: Deep Probabilistic Time Series Imputation. *International Conference on Artificial Intelligence and Statistics*, 1651-1661. https://proceedings.mlr.press/v108/fortuin20a.html

[33] Cao, W., Wang, D., Li, J., Zhou, H., Li, L. and Li, Y. (2018) Brits: Bidirectional Recurrent Imputation for Time Series. In: Bengio, S., Wallach, H., Larochelle, H., Grauman, K., Ce-sa-Bianchi, N. and Garnett, R., Eds., *Advances in Neural Information Processing Systems*, **31**.

[34] Yoon, J., Zame, W.R. and van der Schaar, M. (2019) Estimating Missing Data in Temporal Data Streams Using Multi-Directional Recurrent Neural Networks. *IEEE Transactions on Biomedical Engineering*, **66**, 1477-1490. https://doi.org/10.1109/tbme.2018.2874712

[35] Zhan, T., He, Y., Deng, Y., Li, Z., Du, W. and Wen, Q. (2025) Time Evidence Fusion Network: Multi-Source View in Long-Term Time Series Forecasting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **47**, 11220-11233. https://doi.org/10.1109/tpami.2025.3596905

[36] Wang, S., Li, J., Shi, X., Ye, Z., Mo, B., Lin, W., Ju, S., Chu, Z. and Jin, M. (2024) Timemixer++: A General Time Series Pattern Machine for Universal Predictive Analysis. arXiv:2410.16032.