

Article

Deep Learning for Demand Forecasting: A Framework Incorporating Variational Mode Decomposition and Attention Mechanism

Chunrui Lei ¹, Heng Zhang ^{1,*}, Zhigang Wang ² and Qiang Miao ¹

¹ College of Electrical Engineering, Sichuan University, Chengdu 610065, China; 2022223035096@stu.scu.edu.cn (C.L.); mqiang@scu.edu.cn (Q.M.)

² Jiangsu Sinoclouds S&T Co., Ltd., Zhenjiang 212000, China; lmfever@163.com

* Correspondence: hengzhang27@scu.edu.cn

Abstract: Accurate demand forecasting is crucial for modern supply chain management, forming the foundation for inventory optimization, cost control, and service level improvement. However, demand time series data often exhibit high volatility and diverse patterns, further complicated by the rapid expansion and heterogeneity of data sources. These challenges can result in significant degradation in predictive accuracy when traditional models are applied to complex demand datasets. To address these challenges, this study proposes an end-to-end demand forecasting framework leveraging Variational Mode Decomposition (VMD) and attention mechanisms. The framework first employs VMD to decompose raw demand time series into multiple modes to extract hierarchical features, including trends, seasonal patterns, and short-term variations. Subsequently, an attention mechanism is introduced to dynamically capture and integrate demand sequences alongside contextual information, enhancing the focus on critical features and improving predictive performance. Experimental results demonstrate that the proposed method achieves superior predictive accuracy compared to conventional approaches, with a 37% reduction in Mean Absolute Error (MAE) relative to baseline models. This substantial improvement in demand forecasting accuracy provides actionable insights for decision-makers, enabling more efficient inventory control, production planning, and overall supply chain optimization.



Academic Editor: Abraham Kabutey

Received: 19 December 2024

Revised: 9 February 2025

Accepted: 18 February 2025

Published: 19 February 2025

Citation: Lei, C.; Zhang, H.; Wang, Z.; Miao, Q. Deep Learning for Demand Forecasting: A Framework Incorporating Variational Mode Decomposition and Attention Mechanism. *Processes* **2025**, *13*, 594. <https://doi.org/10.3390/pr13020594>

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

Keywords: demand forecasting; variational mode decomposition; attention mechanism; deep learning

1. Introduction

With the rapid development of globalization and informatization, the complexity of supply chains has been continuously increasing. Enterprises are not only expected to adapt to the dynamically changing market demands but also to coordinate resource allocation and operational efficiency across upstream and downstream segments. Against this backdrop, the significance of supply chain management has become more pronounced. In modern economic systems, supply chain management functions as a critical link connecting production, distribution, and consumption, directly influencing a company's market competitiveness and economic performance [1]. Through effective supply chain management, businesses can optimize resource allocation, reduce inventory costs, enhance logistics efficiency, and swiftly adapt to market fluctuations [2], thereby securing a competitive advantage in an intensely competitive environment [3].

Among the various components of supply chain management, demand forecasting is widely recognized as a pivotal element of the supply chain, directly influencing critical decisions such as procurement planning [4], production scheduling, inventory management, and logistics distribution [5]. Accurate demand forecasting not only helps businesses minimize the risks of stockouts due to under-forecasting and inventory surpluses from over-forecasting [6], but also facilitates lean operations in a dynamic market environment, thereby strengthening the overall resilience of the supply chain [7]. In today's market landscape, enterprises face numerous challenges, including demand diversification, heightened volatility, and a diverse range of product categories and supply sources [8]. Consequently, improving the accuracy and reliability of demand forecasting through advanced technological methods has emerged as a key focus and a persistent challenge in supply chain management research [9].

Substantial advancements have been achieved in the research and application of supply chain demand forecasting [10]. Traditional statistical methods, such as ARIMA [11] and exponential smoothing [12], are noted for their simplicity and interpretability. These methods primarily rely on analyzing autocorrelation in historical data to extrapolate future trends based on past patterns. However, they often overlook the influence of external factors on product demand and lack the capacity to capture random fluctuations within demand sequences. To address these limitations, models like ARIMAX [13] and ETSX [8] have been developed, integrating autocorrelation with external conditions, resulting in moderate improvements in forecasting performance. As market demand becomes increasingly diverse and volatile, demand time series frequently exhibit pronounced non-stationarity and intricate nonlinear dynamics, rendering traditional methods based on stationarity assumptions inadequate for modeling such complexities. Additionally, the expansion of product varieties and sources has significantly increased the scale of demand data, further exposing the shortcomings of traditional approaches in big data environments. Machine learning and deep learning methods, with their robust nonlinear modeling capabilities, dynamic adaptability, and ability to process multimodal, large-scale data, have emerged as pivotal tools in contemporary demand forecasting research [14]. Recurrent Neural Networks (RNNs) [15], owing to their recursive structures, effectively capture short-term dependencies in demand sequences, thereby enhancing prediction accuracy. However, the limitations of traditional RNNs in handling long sequences have led to the widespread adoption of LSTM [16] and GRU [17], which are adept at modeling medium- to long-term trends in demand patterns. For large-scale datasets, the Transformer model [18], with its parallel computing capabilities and flexible temporal correlation mechanisms, has demonstrated promising performance in demand forecasting. Moreover, hybrid models that combine statistical methods with deep learning [19] or integrate machine learning and deep learning approaches [20] are gaining traction, harnessing their complementary strengths to improve forecasting accuracy. Despite these advancements, sales demand sequences often encompass multiple modes, including trends, seasonality, and short-term fluctuations. Existing methods typically analyze raw sequences directly, failing to effectively decompose and utilize these complex modal characteristics, thereby limiting prediction accuracy. Furthermore, few studies adequately assess the influence of critical features when integrating demand sequence characteristics with conditional information. This oversight can lead to information redundancy or loss, ultimately undermining forecasting performance.

Building upon these challenges, this study proposes an end-to-end demand forecasting framework based on Variational Mode Decomposition (VMD) and attention mechanisms. The framework aims to enhance the model's accuracy and robustness in complex demand forecasting tasks by decomposing the raw time series, reconstructing them, and dynamically mining and integrating critical features. This approach facilitates informed

decision-making in modern supply chain management. The key contributions of this study are outlined below:

- (1) The use of VMD to decompose the raw demand sequence into multiple intrinsic mode functions extracts essential components, including trends, periodicity, and high-frequency fluctuations. Additionally, a self-attention mechanism is introduced to dynamically model the impact of conditional factors, focusing on critical time-step features and capturing the complex nonlinear dynamic relationships between external factors and demand. These two components together form a comprehensive data reconstruction module, ensuring the simultaneous integration of historical sales sequences and conditional information, thereby improving forecasting performance.
- (2) In terms of feature fusion, this study employs an attention mechanism to adaptively assign weights to each modal and conditional feature, further enhancing the representation of key features. This approach significantly enhances the accuracy and robustness of the forecasts.
- (3) Through experimental analysis, the proposed framework outperforms conventional methods across multiple evaluation metrics, achieving a 37% reduction in the MAE compared to the baseline BiGRU model. This significant improvement enhances forecasting performance, offering more reliable insights for supply chain management.

The structure of the paper is as follows: Section 2 provides a comprehensive review of the relevant literature on demand forecasting. Section 3 presents a thorough explanation of the proposed deep learning-based demand forecasting framework. Section 4 presents a comprehensive experimental analysis, including comparative experiments and model structure ablation studies. Finally, Section 5 concludes the paper, summarizing the research findings and providing suggestions for future research directions.

2. Related Work

Demand forecasting, as a core component of supply chain management, primarily focuses on accurately predicting future demand using historical data and various influencing factors, which in turn optimizes inventory control, production planning, and logistics distribution. However, due to the inherent characteristics of demand data, such as non-linearity, non-stationarity, large volumes, and complex external factors, the development of efficient and robust forecasting models remains a central focus of ongoing research. Demand forecasting methods are generally classified into three categories: traditional statistical methods, conventional machine learning methods, and deep learning methods. Traditional methods primarily rely on linear assumptions and time series features, which are effective in demand scenarios exhibiting strong regularity but are less effective in handling complex nonlinear problems. Machine learning methods, by introducing nonlinear modeling capabilities, enhance the flexibility of forecasts, though they require extensive feature engineering. Deep learning further overcomes the limitations of feature design, leveraging its powerful automatic feature extraction and modeling capabilities, and has gradually become the mainstream technique for solving complex demand forecasting problems. Therefore, the following discussion will outline the research progress in the field of demand forecasting, summarizing the relevant works and their limitations from three distinct perspectives.

Traditional methods primarily rely on statistical and time series analysis methodologies, analyzing historical data and fitting models to predict future demand based on existing data patterns. Dey et al. [21] developed and compared various linear and nonlinear regression models, along with the ARIMA model, for gasoline demand forecasting, achieving satisfactory forecasting results. Jamil et al. [22] used the ARIMA model to predict future hydropower demand based on historical data and compared the results with actual

values, validating the accuracy and effectiveness of the forecast. Huang et al. [23] designed a forecasting model that integrates ARIMA and adaptive filtering methods for predicting demand in medical services, enhancing the accuracy of forecasts. Pereira et al. [24] studied the autoregressive integrated moving average with external input (ARIMAX) model, demonstrating its superiority in short-term demand forecasting. While statistical methods provide a foundation for demand forecasting, their limitations become increasingly evident as market conditions evolve and the complexity of data characteristics increases. Particularly when confronted with nonlinear relationships, large-scale data, and the impact of complex external factors, traditional statistical methods often struggle to address these challenges effectively.

Compared to traditional statistical methods, machine learning methods do not require strict assumptions of linearity or stationarity, providing enhanced nonlinear modeling capabilities and the ability to manage high-dimensional features effectively. These methods can integrate external influencing factors and contextual information, enhancing the accuracy and robustness of demand forecasting. Ağbulut et al. [25] used Support Vector Machines (SVM) and artificial neural networks to forecast CO₂ emissions and energy demand based on traffic, with the goal of reducing energy consumption in the transportation sector. Javanmard et al. [26] applied six machine learning algorithms to predict energy demand in a certain region until 2040, optimizing different integrated sectors using two optimization algorithms, substantially improving forecasting accuracy. Guo et al. [27] proposed a hybrid model combining Prophet and Support Vector Regression (SVR) to analyze demand in a manufacturing industry, demonstrating the model's accuracy and robustness through empirical validation. Rao et al. [12] first used random forests to identify the driving factors of energy demand, established an SVR-based model to predict multiple types of energy demand, and conducted an in-depth analysis of the forecasting results in real-world contexts. Huber et al. [28] evaluated several machine learning methods, including artificial neural networks and gradient-boosted decision trees, for demand forecasting in retail. They found that machine learning methods yielded strong prediction results and were particularly well-suited for large-scale demand forecasting applications.

With the rapid growth in the scale and complexity of demand data, the limitations of machine learning methods become increasingly evident. Deep learning, by constructing multi-layer neural networks, can automatically learn features extracted from raw data and model complex nonlinear relationships through deep feature abstraction. In processing complex demand data, capturing both long- and short-term dependencies in time series, and uncovering latent dynamic patterns, deep learning has significant advantages over traditional statistical and machine learning methods. Punia et al. [29] proposed a forecasting method combining LSTM and random forests to model complex relationships between time and regression types, and validated the approach using real multivariate datasets from multichannel retailers. Nikolopoulos et al. [30] developed a cross-time forecasting framework based on LSTM, demonstrating its superior forecasting performance through various performance metrics and statistical analyses. Vallés-Pérez et al. [18] utilized deep learning techniques such as Transformer to predict daily sales at customer, store, and product levels using a simple sequence-to-sequence architecture, with validation conducted using Kaggle competition data. Ma et al. [31] designed a convolutional neural network-based multi-branch meta-learner, adapting the weights of base predictors, thereby significantly enhancing forecasting accuracy. Feng et al. [32] developed a federated model-based meta-learning prediction method that not only achieves high accuracy but also demonstrates strong robustness against data heterogeneity and packet loss. Furthermore, Zhu et al. [33] proposed a graph neural network prediction framework based on a novel dynamic graph structure, which outperforms state-of-the-art methods and exhibits superior accuracy.

However, deep learning models often require large volumes of training data to learn intricate feature patterns. In scenarios with limited data, the model may fail to fully learn the features, thereby impairing forecasting performance. Moreover, most researchers often input all features directly into the model for prediction analysis. However, the contributions of various external influencing factors and other features to the prediction are not uniform. Without assessing the impact of key features, this can result in redundancy or loss of information, ultimately compromising the model's performance.

3. Methodology

3.1. Problem Description

In this study, the objective is to predict the sales volume of a given product over a future period of length H , based on sales data from the past w time units prior to time t . In addition to the product's historical sales data, other influencing factors, such as temperature and promotions, are also incorporated to assist in the prediction. Therefore, the forecasting problem is formulated as shown in Equation (1):

$$\hat{y}_{t+1:t+H} = f(y_{t-w:t}, x_{1:i,t-w:t}) \quad (1)$$

where, $\hat{y}_{t+1:t+H}$ denotes the sales volume of this commodity from time $t + 1$ to time $t + H$, which is the prediction target of this study. The independent variable $y_{t-w:t}$ is the historical sales volume of the item from time $t-w$ to t , where w is the length of the sliding window. $x_{1:i,t-w:t}$ is the historical value of the i th influence factor from time $t-w$ to t . Finally, $f(\cdot)$ is the prediction function, which specifically refers to the deep learning prediction framework based on VMD and attention mechanism proposed in this study.

3.2. Demand Forecasting Based on Variational Mode Decomposition and Dual Attention Mechanism

Considering that existing forecasting models inadequately decompose and fully leverage the modal information in the original sales sequences, as well as evaluate the influence of key conditional factors, this paper introduces an end-to-end deep learning demand forecasting framework utilizing VMD and attention mechanisms to improve prediction accuracy. As depicted in Figure 1, the framework is primarily composed of four components: data preprocessing, data reconstruction via VMD-SA, feature enhancement and fusion using the attention mechanism, and ultimately, model prediction employing BiGRU.

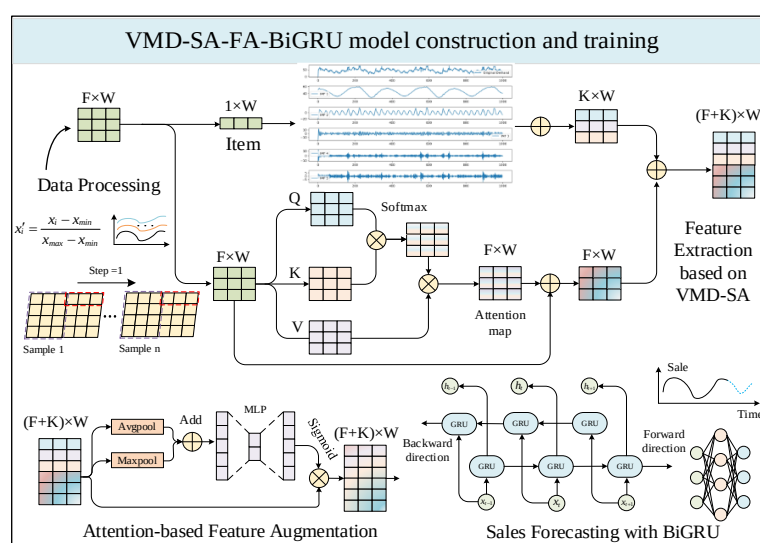


Figure 1. Demand forecasting based on variational mode decomposition and attention mechanism.

3.2.1. Data Preprocessing

In this study, data preprocessing of sales data is a crucial step in ensuring the quality of input features for the model. Raw sales data typically contains multidimensional features from various sources, which may vary in numerical ranges and units of measurement. As a typical time series, sales data demonstrates strong temporal dependencies and trends. Proper data preprocessing is not only directly linked to the model's learning performance but also affects its prediction speed and generalization ability. Therefore, two preprocessing techniques were applied to the raw data: normalization and sliding window operations, to ensure the accuracy of demand forecasting.

(1) Data normalization

Feature data in product sales originates from various sources, including historical sales volume, promotional information, and calendar features, which may differ in numerical ranges and units of measurement. If used directly, the model may be biased toward features with larger numerical ranges, neglecting the importance of other features. To address this, normalization is applied to the data using Equation (2), thereby eliminating differences in units of measurement.

$$x' = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \quad (2)$$

Here, x' represents the data obtained after the normalization of x . $x_{\max}$ and $x_{\min}$ denote the maximum and minimum values of the feature, respectively.

(2) Sliding window

To more effectively capture short-term trends and long-term dependencies in the time series, a sliding window operation is applied to the data. A fixed-size rolling window is applied to the original time series data to extract input features and target labels. As illustrated in Figure 2, the fixed window size is set to W , with a prediction horizon of H and a sliding step of 1.

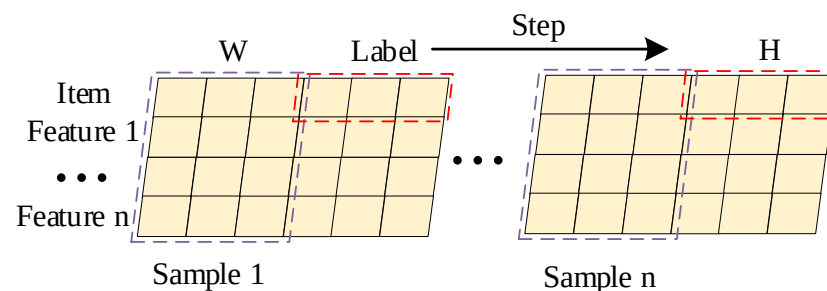


Figure 2. Illustration of the sliding window operation for samples.

3.2.2. Data Reconstruction Based on VMD-SA

(1) Variational mode decomposition (VMD) [34]

VMD is a signal processing technique that decomposes complex non-stationary signals into several relatively stationary intrinsic mode functions. In demand forecasting using sales data, the sales time series often contain various trend and periodic components, such as long-term trends, seasonal fluctuations, and random noise. These different components may interfere with model learning, and directly using the raw data could prevent the model from accurately capturing the underlying patterns. By applying VMD decomposition, we can extract the specific features of each mode, thereby enhancing the accuracy of demand forecasting.

The overall framework of VMD is a variational problem, with the constrained variational problem model as shown in Equation (3):

$$\left\{ \begin{array}{l} \min_{\{u_k\}, \{w_k\}} \left\{ \sum_k^K \left\| \partial_t \left[\left(\delta(t) + \frac{j}{\pi t} \right) * u_k(t) \right] e^{-jw_k t} \right\|_2^2 \right\} \\ \text{s.t.} \sum_k^K u_k(t) = x(t) \end{array} \right\} \quad (3)$$

where, $u_k = \{u_1, u_2, \dots, u_n\}$ and $w_k = \{w_1, w_2, \dots, w_n\}$ are the k th IMF component and its center frequency of the signal $x(t)$ after VMD decomposition. $\delta(t)$ is Dirac delta function; K is the number of decomposed IMF components and ∂_t is the gradient operation.

By introducing the quadratic penalty factor α and Lagrange multiplier λ , the unconstrained augmented Lagrange function is obtained as shown in Equation (4):

$$L(\{u_k\}, \{w_k\}, \lambda) = \alpha \sum_{k=1}^K \left\| \partial_t \left[\left(\delta(t) + \frac{j}{\pi t} \right) * u_k(t) \right] e^{-jw_k t} \right\|_2^2 + \langle \lambda(t), x(t) - \sum_{k=1}^K u_k(t) \rangle \quad (4)$$

ADMM is used to iteratively solve, and to find the optimal solutions of the variational equations. The updated equations are as follows:

$$\hat{u}_k^{n+1}(\omega) = \frac{f(\omega) - \sum_{i \neq k} \hat{u}_i(\omega) - \frac{\hat{\lambda}^n(\omega)}{2}}{1 + 2\alpha(\omega - \omega_k^n)^2} \quad (5)$$

$$w_k^{n+1}(\omega) = \frac{\int_0^\infty \omega |\hat{u}_k^{n+1}(\omega)|_2^2 d\omega}{\int_0^\infty |\hat{u}_k^{n+1}(\omega)|_2^2 d\omega} \quad (6)$$

$$\hat{\lambda}_k^{n+1}(\omega) = \hat{\lambda}^n(\omega) + \tau(\hat{x}(\omega) - \sum_k \hat{u}_k^{n+1}(\omega)) \quad (7)$$

Among them, $\hat{u}_k^{n+1}$, $x(t)$ and λ^n are Fourier-transformed to obtain $\hat{u}^n(\omega)$, $\hat{x}(\omega)$, and $\hat{\lambda}^n(\omega)$, and the condition for terminating the iteration is shown in Equation (8):

$$\sum_k \frac{\left\| \hat{u}_k^{n+1}(\omega) - \hat{u}_k^n(\omega) \right\|_2^2}{\left\| \hat{u}_k^{n+1}(\omega) \right\|_2^2} < \varepsilon \quad (8)$$

(2) Self-attention mechanism with residual structure [35]

The self-attention mechanism can dynamically allocate weights based on the correlation between different time steps, enabling the primary model to focus more effectively on critical time points while suppressing irrelevant information. This enables the model to effectively identify key factors influencing product demand, such as seasonal fluctuations and promotional effects. The core concept is that, for input sequence data, the model calculates the correlation of each position relative to other positions in order to determine its importance within the sequence, as illustrated in Figure 3.

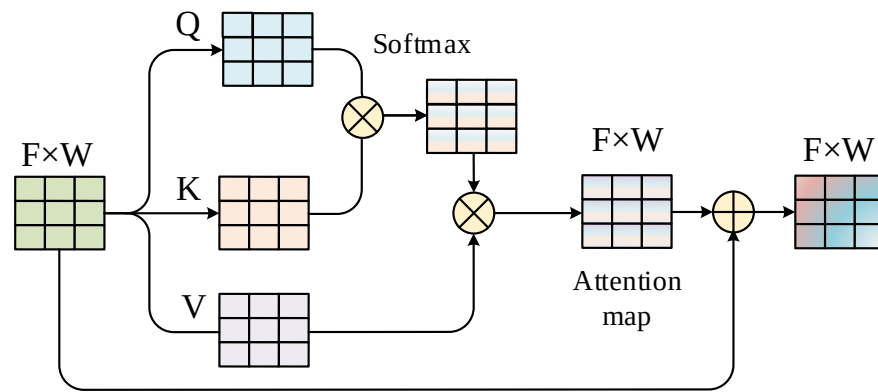


Figure 3. Self-attention mechanism with residual structure.

For the input multivariate time series data $X \in R^{F \times W}$, three linear transformations are performed to map it to the Query, Key, and Value matrices, which are calculated as follows:

$$Q = XW_Q, K = XW_K, V = XW_V \quad (9)$$

where W_Q , W_K and W_V denote the corresponding linear maps, respectively. Q , K and V denote Query, Key, and Value matrices, respectively.

By calculating the similarity between the Query and Key, the attention level of each position relative to other positions is determined. Then, the Softmax function is applied to obtain the attention weights for each position. Finally, the resulting attention weight matrix is multiplied by the Value matrix to obtain the final weighted output representation, as shown in Equation (10).

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right) \cdot V \quad (10)$$

To further improve the stability of the self-attention mechanism, this study incorporates a residual structure, as illustrated in Figure 3. The incorporation of the residual structure allows the mechanism to not only transmit information across layers but also retain the original data. The computational process of the enhanced self-attention mechanism is presented in Equation (11):

$$X_2 = X + Attention(Q, K, V) \quad (11)$$

where, X_2 is the final self-attention output.

Therefore, the final output $X \in R^{(K+F) \times W}$ result of the data reconstruction module is shown in Equation (12):

$$X = concat(X_1, X_2) \quad (12)$$

where, $X_1 \in R^{K \times W}$ is the feature matrix concatenated by IMFs after VMD, K is the number of IMF components, and W is the rolling window size. $X_2 \in R^{F \times W}$ is the final output of self-attention and F is the feature dimension.

3.2.3. Feature Enhancement Fusion Based on Global Attention Mechanism

Using the previous data reconstruction module, we constructed a feature matrix that integrates multimodal information and global time-series dependencies. However, directly utilizing these features may lead to issues such as information redundancy or the loss of critical information. Therefore, in this study, an attention mechanism is employed to refine and optimize the generated feature matrix. This mechanism aims to dynamically assess the importance of features, extract key features, and assign them higher weights, thereby

achieving a deeper fusion of modal features and global influential factors. This refined feature set provides more accurate inputs for subsequent predictive models, as illustrated in Figure 4.

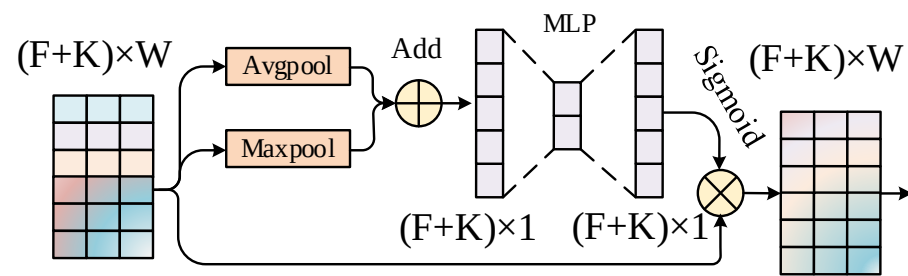


Figure 4. Feature enhancement fusion based on attention mechanism.

(1) Global information embedding

For the feature matrix obtained after data reconstruction, a 1D pooling operation is applied to compress the global information, thereby encoding the multidimensional features. Additionally, to better capture the global information and significant features from different attributes, both max pooling and average pooling are simultaneously used for aggregation. The results of these two pooling operations are then summed to obtain feature values that incorporate both global pooled and averaged features.

(2) Adaptive fusion correction

To better utilize the aggregated feature information from the previously mentioned compression operation, this study employs a simple gating mechanism with a fusion truncation function to model the nonlinear relationships within the feature vectors. Specifically, a network comprising two fully connected layers is used for fusion correction. This process primarily involves a dimensionality reduction fully connected layer with a decay factor of r , a ReLU nonlinear activation function, and a dimensionality expansion fully connected layer. The specific operation is described in Equation (13):

$$X' = \sigma(W_2(\delta(W_1(\text{avg}(X) + \max(X)))))) \cdot X \quad (13)$$

where, $X \in R^{(F+K) \times W}$ is the feature matrix obtained by the data reconstruction module, and $X' \in R^{(F+K) \times W}$ is the weighted feature map obtained by the attention mechanism. σ is the Sigmoid function with a gated mechanism to calculate the feature weights. δ is the ReLU activation function, and W_1 and W_2 are the fully connected layers for raising and reducing dimensions, respectively.

By incorporating the attention mechanism, we can effectively enhance and fuse features from the feature matrix obtained through the data reconstruction module. This is accomplished by dynamically evaluating the importance of each feature and assigning different weights based on its contribution to predictive performance. In the context of sales data, where factors such as varying time scales, seasonal fluctuations, and promotional effects are intertwined, the application of this attention mechanism not only improves feature discriminability and information transfer efficiency but also enhances the model's ability to learn from complex data patterns. This, in turn, provides more accurate inputs for subsequent predictions.

3.2.4. Demand Forecasting with BiGRU

In the backbone network, we utilize the BiGRU model for demand forecasting. Compared to traditional Recurrent Neural Networks (RNNs), the GRU effectively mitigates the vanishing gradient problem by introducing gating mechanisms, enabling it to better

capture long-term dependencies. Compared with LSTM, GRU simplifies the gating mechanism, reducing the number of parameters and enhancing computational efficiency. At the same time, it maintains a comparable modeling capability in certain tasks, making it well-suited for handling large-scale, high-dimensional, and complex time series data with multiple features [36]. Building on this, BiGRU further extends the capabilities of GRU with its bidirectional structure, allowing the model to capture both forward and backward dependencies within the time series, thereby providing a global understanding of the context at each time step. This bidirectional modeling approach not only strengthens the ability to extract key features but also more comprehensively captures the nonlinear patterns and complex dependencies in sales data. As a result, BiGRU offers significant advantages in capturing global time-series features and improving predictive performance.

The Gated Recurrent Unit (GRU) [36] is an improved version of the RNN, as illustrated in Figure 5. The core of the GRU consists of two gating units: the update gate and the reset gate. The update gate determines how much information from past memory should be retained at the current time step, as shown in Equation (14). The reset gate controls the degree to which the current input is fused with past information, as shown in Equation (15). During computation, the GRU selectively forgets old information and introduces new information via the gating units, dynamically adjusting the update of the hidden state, as shown in Equation (17):

$$z_t = \sigma(W_z x_t + U_z h_{t-1} + b_z) \quad (14)$$

where z_t is the update gate, σ is the Sigmoid activation function, x_t is the current input, and h_{t-1} is the hidden state of the previous time step. W_z , U_z and b_z are the learnable parameters:

$$r_t = \sigma(W_r x_t + U_r h_{t-1} + b_r) \quad (15)$$

where, r_t is the reset gate. W_r , U_r , and b_r are the learnable parameters:

$$\tilde{h}_t = \tanh(W_h x_t + r_t \odot (U_h h_{t-1}) + b_h) \quad (16)$$

where, $\tilde{h}_t$ is the candidate hidden state, which is the combination of the current input and the previous hidden state, and $\odot$ represents the element-wise product:

$$h_t = z_t \odot h_{t-1} + (1 - z_t) \odot \tilde{h}_t \quad (17)$$

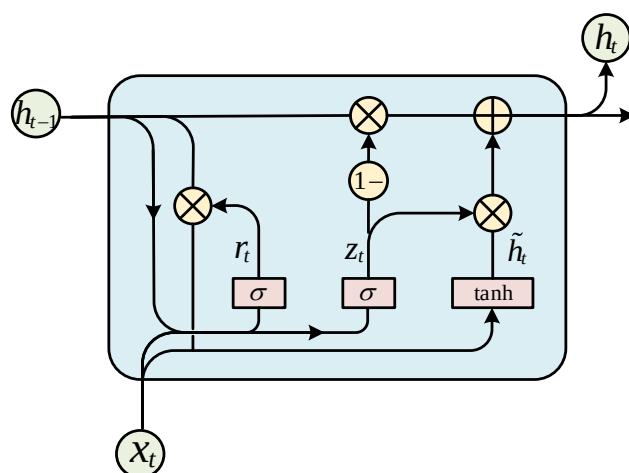


Figure 5. Architectural representation of the gated recurrent unit (GRU).

The Bidirectional Gated Recurrent Unit (BiGRU) introduces a bidirectional structure on top of the GRU, enabling it to simultaneously capture features from both forward and backward directions in the time series, as illustrated in Figure 6. BiGRU consists of two independent GRU units: one unit extracts forward dependency features by processing the time series in its natural order (from past to future), while the other unit extracts backward dependency features by processing in reverse order (from future to past). The outputs of these two GRU units are fused at each time step to form a comprehensive feature representation, providing global contextual information for the current time step. This bidirectional modeling approach of BiGRU is particularly well-suited for handling complex temporal dependencies and asymmetric features, making it especially effective in capturing global information from time series and enhancing prediction accuracy. It is particularly advantageous in tasks such as demand forecasting, where both long-term dependencies and short-term fluctuations are prevalent.

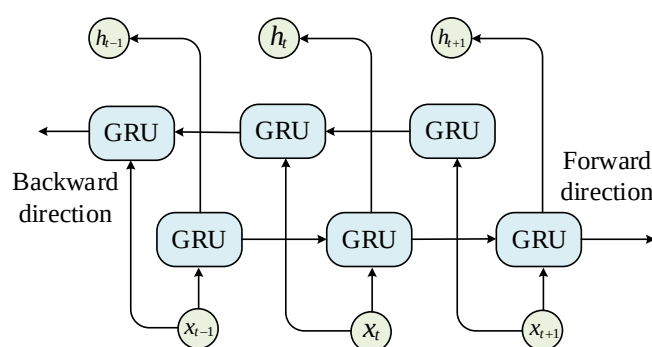


Figure 6. The bidirectional gated recurrent unit.

4. Experimental Analysis

In this section, we first present the dataset used and the prediction evaluation metrics. Subsequently, through experiments, we compare the proposed model with commonly used time series forecasting models to assess its effectiveness. Finally, we perform ablation experiments to evaluate the necessity of the proposed framework modules. All experiments were conducted on a server equipped with an NVIDIA GeForce RTX 4090 GPU, sourced from GIGABYTE Technology, Chengdu, China, running the CentOS Stream 9 operating system.

4.1. Data Description

The dataset used in this study is obtained from a retail company and contains historical sales data from 10 stores spanning the period from 2013 to 2018. Provided in Excel format, the dataset includes daily sales records for four different products across various stores, along with external factors influencing sales, such as holidays, daily temperature, promotional activities, and date-related features. Regarding the data structure, each record consists of the date, store ID, product ID, daily sales volume, and relevant external condition variables. Specifically, date features are presented in the year/month/day format, temperature data represents the local daily average temperature in degrees Celsius, and promotional information is normalized to a range of 0 to 1 to indicate the intensity of promotions over different time periods. Additionally, holiday information is encoded as a binary variable, enabling the model to capture the impact of special dates on sales performance.

The primary objective of this study is to predict sales volumes within a given forecasting horizon based on historical sales data and external contextual factors. The incorporation of these conditional factors enhances the model's capability to comprehend the broader sales environment, enabling it to more accurately capture fluctuations in product demand.

4.2. Evaluation Index

In demand forecasting, the accuracy of predictions has a direct impact on supply chain decisions, including inventory management and operational cost control. To comprehensively evaluate the performance of the forecasting model, this study employs three widely used evaluation metrics: Mean Squared Error (MSE), Mean Absolute Error (MAE), and Mean Absolute Percentage Error (MAPE). These metrics assess the deviation between the model's predicted values and the actual sales, offering a comprehensive evaluation of the model's ability to capture demand fluctuations from multiple perspectives.

(1) Mean Squared Error (MSE)

MSE is one of the most widely used regression evaluation metrics, quantifying the average squared error between predicted and actual values. The equation for MSE is given as follows. MSE effectively captures the model's ability to respond to significant demand fluctuations, thereby mitigating the risk of inventory shortages or surpluses due to large demand variations:

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (y_i - \hat{y}_i)^2 \quad (18)$$

where, y_i is the true value, $\hat{y}_i$ is the predicted value and n is the number of samples.

(2) Mean Absolute Error (MAE)

MAE primarily calculates the average of the absolute errors between the predicted and actual values, as shown in the following equation. It provides an intuitive assessment of the overall prediction accuracy of the model, helping to determine whether the model meets the accuracy requirements for daily operational decision-making:

$$\text{MAE} = \frac{1}{n} \sum_{i=1}^n |y_i - \hat{y}_i| \quad (19)$$

(3) Mean Absolute Percentage Error (MAPE)

MAPE is used to evaluate the proportion of the forecast error relative to the actual value, as shown in the following equation. By normalizing the error as a percentage, MAPE mitigates the impact of varying demand scales, facilitating cross-comparisons of forecasting performance across different products or time periods.

$$\text{MAPE} = \frac{100\%}{n} \sum_{i=1}^n \left| \frac{y_i - \hat{y}_i}{y_i} \right| \quad (20)$$

4.3. Comparative Experiment

4.3.1. VMD Parameter Selection

In the data reconstruction module, we employ VMD to decompose the complex sales demand time series into several intrinsic mode functions. The number of modes, denoted as K , in the decomposition is a key hyperparameter in VMD. If K is set too small, the decomposition may fail to capture the complex dynamic characteristics of the signal, resulting in the loss of crucial information. Conversely, if K is set too large, excessive redundant information or noise may be introduced, diminishing the effectiveness of the features. Therefore, selecting an optimal value for K is crucial to ensure that the decomposition results are consistent with the prediction model's requirements.

To examine the impact of the number of modes, K , on demand forecasting performance, we conducted a hyperparameter experiment with K values ranging from 1 to 9. By comparing the model's forecasting performance across different values of K , we analyze the varying decomposition results and reveal the sensitivity of the parameter, ultimately

enhancing the accuracy of the model's predictions. The detailed experimental results are presented in Figure 7.

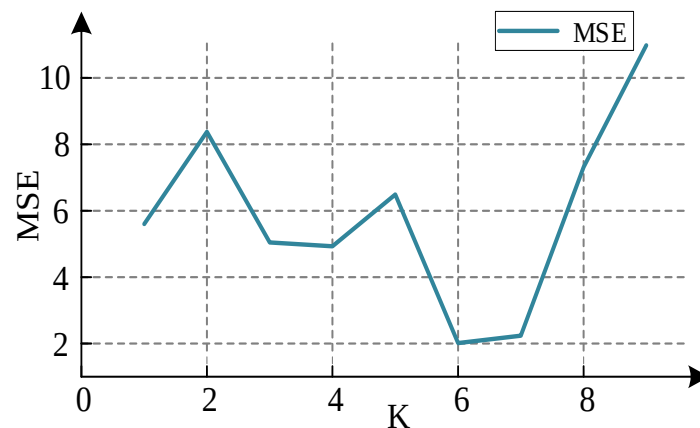


Figure 7. Selection process for variational mode decomposition parameters.

As illustrated in the figure, the model's forecasting performance exhibits variation based on the number of modes. When K is set too small, the number of decomposed modes is insufficient to capture the complex variations in the demand sequence, resulting in the loss of critical features and suboptimal prediction performance. Conversely, when K is set too large, the decomposed modes may include excessive irrelevant information or noise, thereby degrading the quality and expressive power of the features, leading to a reduction in model performance. As shown in the figure, the model's prediction performance is better when K is set to 6 or 7. To reduce computational complexity, K was set to 6.

4.3.2. Comparison of Methods

To comprehensively evaluate the effectiveness of the proposed method, a series of comparative experiments were performed. In this section, we compare the proposed model with seven benchmark models: ARIMA, LSTM, GRU, TCN, BiGRU, Informer [37], and TiDE [38], with the BiGRU model selected as the baseline. The experiments were conducted using Python 3.7 as the programming language. The ARIMA model was implemented and optimized with the Statsmodels 0.14.0 package, while the deep learning models were developed using the PyTorch 1.12.1 framework. The dataset is split into training and testing subsets, with the first 80% used for model training and the remaining 20% reserved for evaluation and performance analysis.

The specific experimental setup and parameter configurations are provided in Table 1. Window size denotes the length of the sliding window, representing the number of historical data points utilized in each training instance to capture temporal dependencies. Prediction length specifies the number of future time steps the model aims to forecast, determining the scope of its output. Batch size specifies the number of samples processed per training update, affecting both the stability of training and computational efficiency. Epoch refers to the number of times the entire training dataset is passed through the neural network, influencing the model's capacity to fit the data. Learning rate controls the step size of gradient updates, influencing both the speed and stability of model convergence. For the optimization process, the Adam optimizer is employed, leveraging momentum and adaptive learning rates to enhance training efficiency and accelerate convergence. The MSE is employed as the loss function to quantify the discrepancy between predicted and actual values. Notably, the three key deep learning hyperparameters—Batch size, Epoch, and Learning rate—are optimized using a grid search strategy to identify the optimal configuration.

Table 1. Experimental parameter configuration.

Experimental Parameters	Parameter Setting
Window size	90
Prediction length	7
Batch size	64
Epoch	200
Learning rate	0.001
Optimizer	Adam
Loss	MSE

The experimental parameters of each comparison model are shown in Table 2.

Table 2. Comparison of model experimental parameter configurations.

Methods	Parameter Setting
ARIMA	p: 2,d: 1,q: 1
LSTM	num_layers: 2 hidden_size: 32
GRU	num_layers: 2 hidden_size: 32
TCN	kernel_size = 2 dropout = 0.2
BiGRU	num_layers: 2hidden_size: 32
Informer	n_head = 8 d_layers = 2 hidden_dim:100
TiDE	encoder_layer num: 4 decoder_layer_num: 4 VMD_K: 6
Proposed	SA head: 4 SA dim_head: 8 Attention_ratio: 3

To ensure the validity of comparative experiments, all models are trained with the same hyperparameter settings shown in Table 1. This consistency allows performance differences to more accurately reflect the inherent effectiveness of each algorithm rather than variations in hyperparameter configurations.

Moreover, to ensure the accuracy and robustness of the experimental results, each experiment was repeated five times, and the final results are reported as the mean and standard deviation, as shown in Table 3.

Table 3. Results of the Comparative Experiment.

Methods/Metrics	MSE	MAE	MAPE (%)	Train Time (s)
ARIMA	19.843 ± 2.578	3.822 ± 0.751	7.65 ± 0.67	10
LSTM	8.403 ± 0.346	1.912 ± 0.041	4.37 ± 0.23	76
GRU	8.755 ± 0.406	1.813 ± 0.062	4.11 ± 0.14	69
TCN	10.675 ± 0.649	2.142 ± 0.095	4.85 ± 0.24	103
BiGRU	7.761 ± 0.198	1.730 ± 0.092	3.96 ± 0.09	101
TiDE	5.228 ± 0.145	1.356 ± 0.026	3.37 ± 0.05	156
Informer	6.370 ± 0.514	1.541 ± 0.109	3.50 ± 0.22	321
Proposed	2.117 ± 0.110	1.147 ± 0.066	2.70 ± 0.16	128

As shown in the table, the traditional ARIMA model performs poorly when applied to nonlinear and complex data, as it cannot handle multivariate features or capture long-term dependencies. Consequently, its prediction accuracy is significantly lower than that of the other models. Both the LSTM and GRU models yield similar performance, effectively

capturing long-term dependencies and adapting to nonlinear variations. The TCN model, which captures long-range dependencies through causal and dilated convolutions, is less adept at modeling global context within the sequence compared to LSTM, resulting in relatively weaker performance. The BiGRU model extends the unidirectional structure of GRU to a bidirectional one, allowing it to extract information from both past and future time steps, leading to improved results for complex time series data. The Informer model, based on transformer architecture, enhances efficiency through a sparse attention mechanism, which marginally improves prediction performance. The TiDE model, which incorporates time decomposition, demonstrates strong adaptability and performs well. The proposed model in this paper, however, utilizes VMD decomposition to decompose the sales sequence, followed by data reconstruction with self-attention mechanisms. Attention mechanisms are then employed to mine key features, enhancing and fusing them to improve prediction accuracy. As a result, the proposed model outperforms all others in terms of performance metrics.

To reflect the differences in computational performance across models, the training times for various models are reported in the table above. As shown, the ARIMA model, with its linear computational characteristics, has the shortest training time. However, due to its limitations in nonlinear modeling, its predictive performance is comparatively lower. In contrast, GRU requires fewer parameters than LSTM, resulting in slightly shorter training times, although both are still outperformed by the TCN in terms of training efficiency. The Informer and TiDE network models possess a larger number of parameters, and while they deliver better prediction accuracy compared to standard models, their training times are significantly longer. The network structure proposed in this study incorporates VMD, which increases computational overhead to a certain degree, leading to a longer training time compared to basic models. However, because of the incorporation of attention mechanisms with fewer parameters in other parts of the architecture, the overall training time remains within an acceptable range. Notably, the training time is substantially reduced compared to the Transformer-based Informer model, resulting in better computational performance. This improved efficiency may assist decision-makers in quickly optimizing the supply chain based on the prediction outcomes.

To facilitate a more intuitive comparison of the prediction results, we plot the predicted values from the top three models, as shown in Figure 8.

As shown in Figure 8, the upper section compares the predicted values and actual values for three models. It is clear that the proposed model's predictions are closer to the actual values compared to the other two models. The prediction error of the proposed model is consistently lower than that of traditional models across most periods, particularly during times of significant demand fluctuations. The proposed model is more capable of accurately capturing sharp increases or decreases in sales, while traditional models are less responsive in the short term, leading to larger prediction errors. Importantly, due to the use of Variational Mode Decomposition (VMD) to decompose the original demand sequence and the application of a self-attention mechanism to extract time-dependent information based on conditional factors, the proposed model performs particularly well at points of sudden promotional changes. By contrast, other models exhibit slower responses to demand surges following the start of a promotion and demonstrate a less stable recovery after the promotion ends. Through comprehensive analysis of promotional points, supply chain managers can leverage this model for precise inventory forecasting, thus avoiding inventory shortages caused by unforeseen demand spikes due to promotions. Additionally, the model is effective in predicting the demand drop-off trend after promotions, aiding decision-makers in inventory clearance and subsequent production planning.

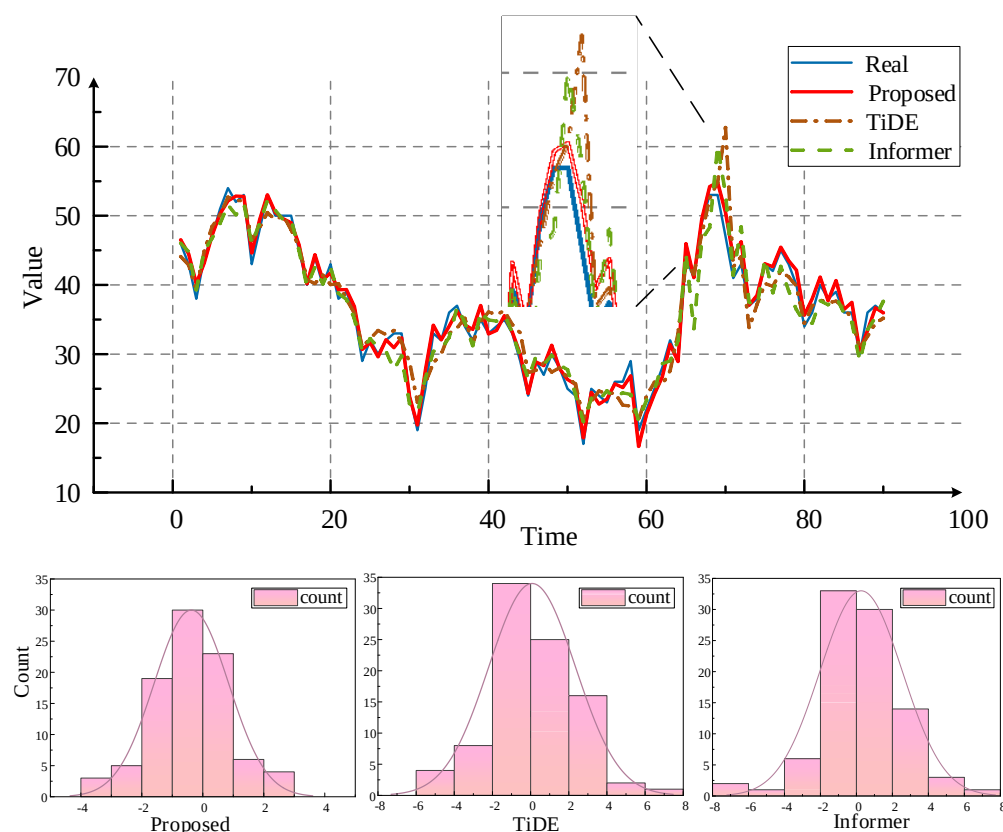


Figure 8. Prediction results and distribution histograms.

To further illustrate the differences in prediction performance, we have plotted the residual distribution histograms of the three models relative to the actual values. Through this visualization, we can visually assess the prediction error distribution of each model and analyze their stability and accuracy. The residuals of all three models are concentrated around zero, with most of the errors falling within a small range, indicating good prediction performance overall. However, the error distribution of traditional models is more dispersed, exhibiting greater fluctuations. Particularly during periods of significant demand variability, such as promotional periods and seasonal changes, the errors of these models become more pronounced, indicated by longer tails. This suggests that traditional models are prone to generating larger prediction errors when dealing with demand fluctuations or strong external influences. By comparison, the residuals of the proposed model are concentrated within the range of $(-4, 4)$, whereas the residuals of the other two models are within $(-8, 8)$. This demonstrates that the proposed model can predict demand with greater accuracy and stability during most periods, exhibiting smaller and more consistent errors. This advantage allows the model to effectively track actual sales fluctuations, providing reliable decision support for supply chain management and aiding in the optimization of inventory management, production scheduling, and other critical operations.

4.4. Ablation Experiment

To validate the effectiveness of the model architecture proposed in this paper, we conduct a series of ablation experiments in this section. By analyzing module removal and replacement, we comprehensively evaluate the performance of each individual module. This in-depth investigation explores the contribution and collaborative effects of each module within the task, further validating the predictive performance of the model.

4.4.1. Data Reconstruction Module

The data reconstruction module is a crucial element of the proposed forecasting framework. It conducts a structured analysis of time-series signals through VMD decomposition and utilizes the self-attention mechanism to capture dynamic relationships among multiple features. To validate the independent contributions and collaborative effects of these two components, we replace each component accordingly while keeping the other modules fixed. By comparing the prediction performance of different models, we seek to emphasize the necessity of the composite feature reconstruction module design. The specific experimental results are shown in Table 4.

Table 4. The results of ablation experiment on the data reconstruction module.

Methods/Metrics	MSE	MAE	MAPE (%)
LSTM	8.40	1.912	4.37
GRU	8.755	1.813	4.11
TCN	10.675	2.142	4.85
BiGRU	7.761	1.730	3.96
VMD-GRU	6.658	1.814	4.17
VMD-LSTM	7.296	1.983	4.51
VMD-TCN	7.061	1.931	4.42
VMD-BiGRU	5.973	1.675	3.89
SA-GRU	7.542	1.624	3.92
SA-LSTM	6.254	1.535	3.52
SA-TCN	8.812	1.843	4.20
SA-BiGRU	6.912	2.068	2.93
VMD-SA-GRU	3.900	1.513	3.45
VMD-SA-LSTM	3.976	1.433	3.34
VMD-SA-TCN	4.268	1.442	3.32
Proposed	2.117	1.147	2.70

First, the VMD decomposition module is placed at the front of each base model, and the decomposed modes are input as features. As observed, the VMD decomposition breaks down the original time-series signal into multiple modal components, allowing the model to focus on capturing the characteristics of each mode. This approach enhances prediction performance to some degree. Furthermore, regarding conditional factors, we use the self-attention mechanism to enhance the model's ability to capture long-range dependencies. By assigning different weights to each time step based on the correlation of historical data, prediction accuracy is significantly improved. Finally, by integrating both components into the data reconstruction module and analyzing their synergistic effects, we observe that the model exhibits superior performance.

4.4.2. Feature Enhancement Fusion Based on GLOBAL Attention Mechanism

Considering the varying importance of different modes and the influence of conditional factors, a global attention mechanism is introduced to focus on key features. To verify the effectiveness of the designed global attention mechanism, a comparative analysis is performed by replacing this part while keeping the other model components fixed. Four comparison schemes are designed in this section: the first involves directly concatenating the feature matrix; the second uses a fully connected layer for concatenation mapping. Finally, the module is substituted with two commonly used attention mechanisms: channel attention and self-attention. The specific experimental results are shown in Table 5.

Table 5. The results of ablation experiment on the feature enhancement fusion.

Methods/Metrics	MSE	MAE	MAPE (%)
Concat	3.365	1.358	3.20
Full connected	4.346	1.542	3.43
SE	2.524	1.265	2.82
SA	3.097	1.304	3.03
Proposed	2.117	1.147	2.703

As demonstrated in Table 5, the complete model with the global attention mechanism outperforms other feature fusion strategies in terms of performance metrics. Compared to direct concatenation, the global attention mechanism is more effective in focusing on key modes and conditional influencing factors, thus laying a solid foundation for subsequent predictions.

4.4.3. Backbone Network

The model employs the BiGRU network as the backbone for prediction to perform time-series forecasting analysis. To validate its effectiveness, this section substitutes the backbone network with LSTM, GRU, and TCN networks. While maintaining consistency in the feature reconstruction and feature fusion modules, the modeling ability of different backbone networks for the supply chain sales sequence is analyzed. The specific experimental results are shown in Table 6.

Table 6. The results of ablation experiment on the backbone network.

Methods/Metrics	MSE	MAE	MAPE (%)
LSTM	3.976	1.433	3.3
GRU	3.900	1.513	3.45
TCN	4.268	1.442	3.32
Proposed	2.117	1.147	2.703

Experimental results indicate that, within this prediction framework, the use of the BiGRU model as the backbone for time-series modeling offers several advantages. When compared to the unidirectional GRU network, its bidirectional time-series modeling ability effectively captures the implicit correlations between historical and potential future patterns in supply chain data. While the TCN can model long-term dependencies, its dynamic response capability to sudden events in non-stationary sequences is relatively weaker. This further demonstrates that the BiGRU network, by integrating bidirectional temporal features, exhibits greater robustness in typical supply chain scenarios such as complex promotions and seasonal fluctuations.

Although the model has achieved good performance in multiple experiments, there are still several limitations. First, the model's performance could degrade when the data quality is poor or when there is a significant amount of missing data. Despite normalization and sliding window processing during the data preprocessing phase, handling extreme values or missing data may still influence the model's prediction accuracy. Second, while the model uses self-attention mechanisms and the BiGRU network to capture dependencies in the time series, its ability to respond to complex external environmental changes (e.g., sudden events or supply chain disruptions) is limited. This could result in suboptimal prediction performance in certain exceptional scenarios. Moreover, the model relies heavily on large amounts of historical data for training, which may result in inaccurate predictions in rapidly changing markets or with limited product data. Future work should focus on enhancing the model's generalization ability, enabling it to perform reliably across various supply chain scenarios and environments (e.g., new product forecasting). Additionally,

incorporating reinforcement learning and other methods to explore the possibility of dynamically adjusting forecasting strategies is expected to be an important direction for future research.

5. Conclusions

With the intensification of global market competition and the increasing complexity of supply chain management, accurate demand forecasting has become essential for businesses to optimize inventory, reduce costs, and improve service levels. However, existing methods often suffer from insufficient modeling of complex demand patterns, neglecting key features, and inconsistent prediction accuracy and robustness. To address these issues, this study proposes a demand forecasting framework that integrates VMD and attention mechanisms. The method first decomposes the original time series using VMD to extract multi-layered feature information, such as trends, seasonality, and short-term fluctuations, thereby capturing the complex patterns in demand data. Next, an attention mechanism is introduced to dynamically weight different time steps and features, emphasizing the significance of key features and effectively avoiding information redundancy and loss. This improves the overall prediction accuracy and robustness. Experimental results show that the proposed framework outperforms conventional methods on multiple evaluation metrics, achieving a 37% reduction in MAE compared to the benchmark model, BiGRU. In contrast, the framework performs better in managing complex demand fluctuations and various types of conditional information, providing more precise data support for decision-making in supply chain management, especially in inventory management, production scheduling, and resource allocation.

Although this study has made significant improvements in forecasting accuracy, challenges remain in handling more complex and variable demand data. Future research should explore more advanced feature selection and fusion methods, as well as enhance the model's real-time performance and computational efficiency. Additionally, incorporating richer external information, such as market trends and macroeconomic data, is expected to further enhance the accuracy and reliability of demand forecasting, thus driving the continued application and development of demand forecasting technology in supply chain management.

Author Contributions: Conceptualization, C.L.; methodology, C.L.; software, C.L.; validation, C.L.; formal analysis, C.L.; investigation, Q.M. and Z.W.; resources, Z.W.; data curation, C.L.; writing original draft preparation, C.L.; writing review and editing, H.Z.; visualization, C.L.; supervision, Q.M. All authors have read and agreed to the published version of the manuscript.

Funding: The research was partially supported by the National Key R&D Program of China (No. 2021YFB3300800, No. 2021YFB3300801 and No. 2021YFB3300803).

Data Availability Statement: Due to privacy or ethical restrictions, the data supporting the reported results cannot be publicly shared.

Conflicts of Interest: Author Zhigang Wang was employed by the Jiangsu Sinoclouds S&T Co., Ltd. The remaining authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

References

1. Khan, S.A.R.; Yu, Z.; Golpira, H.; Sharif, A.; Mardani, A. A state-of-the-art review and meta-analysis on sustainable supply chain management: Future research directions. *J. Clean. Prod.* **2021**, *278*, 123357. [[CrossRef](#)]
2. Mohamadi, N.; Niaki, S.T.A.; Taher, M.; Shavandi, A. An application of deep reinforcement learning and vendor-managed inventory in perishable supply chain management. *Eng. Appl. Artif. Intell.* **2004**, *127*, 107403. [[CrossRef](#)]

3. Lei, C.; Zhang, H.; Yan, X.; Miao, Q. Green Supply Chain Optimization Based on Two-Stage Heuristic Algorithm. *Processes* **2024**, *12*, 1127. [\[CrossRef\]](#)
4. Nia, A.R.; Awasthi, A.; Bhuiyan, N. Industry 4.0 and demand forecasting of the energy supply chain: A literature review. *Comput. Ind. Eng.* **2021**, *154*, 107128.
5. Seyedan, M.; Mafakheri, F. Predictive big data analytics for supply chain demand forecasting: Methods, applications, and research opportunities. *J. Big Data* **2020**, *7*, 53. [\[CrossRef\]](#)
6. Fahimnia, B.; Tan, T.; Tahirov, N. Service-level anchoring in demand forecasting: The moderating impact of retail promotions and product perishability. *Int. J. Forecast.* **2024**, *in press*. [\[CrossRef\]](#)
7. Kuo, T.C.; Chen, K.J.; Shiang, W.J.; Huang, P.B.; Otieno, W.; Chiu, M.C. A collaborative data-driven analytics of material resource management in smart supply chain by using a hybrid Industry 3.5 strategy. *Resour. Conserv. Recycl.* **2021**, *164*, 105160. [\[CrossRef\]](#)
8. Abolghasemi, M.; Beh, E.; Tarr, G.; Gerlach, R. Demand forecasting in supply chain: The impact of demand volatility in the presence of promotion. *Comput. Ind. Eng.* **2020**, *142*, 106380. [\[CrossRef\]](#)
9. Lei, C.; Zhang, H.; Wang, Z.; Miao, Q. Multi-Model Fusion Demand Forecasting Framework Based on Attention Mechanism. *Processes* **2024**, *12*, 2612. [\[CrossRef\]](#)
10. Niu, T.; Zhang, H.; Yan, X.; Miao, Q. Intricate Supply Chain Demand Forecasting Based on Graph Convolution Network. *Sustainability* **2024**, *16*, 9608. [\[CrossRef\]](#)
11. Wang, C.C.; Chien, C.H.; Trappey, A.J. On the application of ARIMA and LSTM to predict order demand based on short lead time and on-time delivery requirements. *Processes* **2021**, *9*, 1157. [\[CrossRef\]](#)
12. Rao, C.; Zhang, Y.; Wen, J.; Xiao, X.; Goh, M. Energy demand forecasting in China: A support vector regression-compositional data second exponential smoothing model. *Energy* **2023**, *263*, 125955. [\[CrossRef\]](#)
13. Falatouri, T.; Darbanian, F.; Brandtner, P.; Udokwu, C. Predictive analytics for demand forecasting—A comparison of SARIMA and LSTM in retail SCM. *Procedia Comput. Sci.* **2022**, *200*, 993–1003. [\[CrossRef\]](#)
14. Yan, X.; Zhang, H.; Wang, Z.; Miao, Q. Probabilistic Time Series Forecasting Based on Similar Segment Importance in the Process Industry. *Processes* **2024**, *12*, 2700. [\[CrossRef\]](#)
15. Sahu, M.K. Advanced AI Techniques for Optimizing Inventory Management and Demand Forecasting in Retail Supply Chains. *J. Bioinform. Artif. Intell.* **2021**, *1*, 190–224.
16. Nguyen, H.D.; Tran, K.P.; Thomassey, S.; Hamad, M. Forecasting and Anomaly Detection approaches using LSTM and LSTM Autoencoder techniques with the applications in supply chain management. *Int. J. Inf. Manag.* **2021**, *57*, 102282. [\[CrossRef\]](#)
17. Peng, T.; Gan, M.; Ou, Q.; Yang, X.; Wei, L.; Ler, H.R.; Yu, H. Railway cold chain freight demand forecasting with graph neural networks: A novel GraphARMA-GRU model. *Expert Syst. Appl.* **2024**, *255*, 124693. [\[CrossRef\]](#)
18. Vallés-Pérez, I.; Soria-Olivas, E.; Martínez-Sober, M.; Serrano-López, A.J.; Gómez-Sanchís, J.; Mateo, F. Approaching sales forecasting using recurrent neural networks and transformers. *Expert Syst. Appl.* **2022**, *201*, 116993. [\[CrossRef\]](#)
19. Li, Y.; Yang, Y.; Zhu, K.; Zhang, J. Clothing sale forecasting by a composite GRU-Prophet model with an attention mechanism. *IEEE Trans. Ind. Inform.* **2021**, *17*, 8335–8344. [\[CrossRef\]](#)
20. Punia, S.; Shankar, S. Predictive analytics for demand forecasting: A deep learning-based decision support system. *Knowl. Based Syst.* **2022**, *258*, 109956. [\[CrossRef\]](#)
21. Dey, B.; Roy, B.; Datta, S.; Ustun, T.S. Forecasting ethanol demand in India to meet future blending targets: A comparison of ARIMA and various regression models. *Energy Rep.* **2023**, *9*, 411–418. [\[CrossRef\]](#)
22. Jamil, R. Hydroelectricity consumption forecast for Pakistan using ARIMA modeling and supply-demand analysis for the year 2030. *Renew. Energy* **2020**, *154*, 1–10. [\[CrossRef\]](#)
23. Huang, Y.; Xu, C.; Ji, M.; Xiang, W.; He, D. Medical service demand forecasting using a hybrid model based on ARIMA and self-adaptive filtering method. *BMC Med. Inform. Decis. Mak.* **2020**, *20*, 237. [\[CrossRef\]](#)
24. Gonçalves, J.N.; Cortez, P.; Carvalho, M.S.; Frazao, N.M. A multivariate approach for multi-step demand forecasting in assembly industries: Empirical evidence from an automotive supply chain. *Decis. Support Syst.* **2021**, *142*, 113452. [\[CrossRef\]](#)
25. Ağbulut, Ü. Forecasting of transportation-related energy demand and CO₂ emissions in Turkey with different machine learning algorithms. *Sustain. Prod. Consum.* **2022**, *29*, 141–157. [\[CrossRef\]](#)
26. Javanmard, M.E.; Ghaderi, S.F. Energy demand forecasting in seven sectors by an optimization model based on machine learning algorithms. *Sustain. Cities Soc.* **2023**, *95*, 104623. [\[CrossRef\]](#)
27. Guo, L.; Fang, W.; Zhao, Q.; Wang, X. The hybrid PROPHET-SVR approach for forecasting product time series demand with seasonality. *Comput. Ind. Eng.* **2021**, *161*, 107598. [\[CrossRef\]](#)
28. Huber, J.; Stuckenschmidt, H. Daily retail demand forecasting using machine learning with emphasis on calendric special days. *Int. J. Forecast.* **2020**, *36*, 1420–1438. [\[CrossRef\]](#)
29. Punia, S.; Nikolopoulos, K.; Singh, S.P.; Madaan, J.K.; Litsiou, K. Deep learning with long short-term memory networks and random forests for demand forecasting in multi-channel retail. *Int. J. Prod. Res.* **2020**, *58*, 4964–4979. [\[CrossRef\]](#)

30. Punia, S.; Singh, S.P.; Madaan, J.K. A cross-temporal hierarchical framework and deep learning for supply chain forecasting. *Comput. Ind. Eng.* **2020**, *149*, 106796. [[CrossRef](#)]
31. Ma, S.; Fildes, R. Retail sales forecasting with meta-learning. *Eur. J. Oper. Res.* **2021**, *288*, 111–128. [[CrossRef](#)]
32. Feng, C.; Shao, L.; Wang, J.; Zhang, Y.; Wen, F. Short-term Load Forecasting of Distribution Transformer Supply Zones Based on Federated Model-Agnostic Meta Learning. *IEEE Trans. Power Syst.* **2024**, *40*, 31–45. [[CrossRef](#)]
33. Zhu, N.; Wang, Y.; Yuan, K.; Yan, J.; Li, Y.; Zhang, K. GGNet: A novel graph structure for power forecasting in renewable power plants considering temporal lead-lag correlations. *Appl. Energy* **2024**, *364*, 123194. [[CrossRef](#)]
34. Dragomiretskiy, K.; Zosso, D. Variational mode decomposition. *IEEE Trans. Signal Process.* **2013**, *62*, 531–544. [[CrossRef](#)]
35. Chung, J.; Gulcehre, C.; Cho, K.; Bengio, Y. Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv* **2014**, arXiv:1412.3555.
36. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, L.; Polosukhin, I. Attention is all you need. In *Advances in Neural Information Processing Systems, Proceedings of the 31st Annual Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017*; Curran Associates, Inc.: Red Hook, NY, USA, 2017.
37. Zhou, H.; Zhang, S.; Peng, J.; Zhang, S.; Li, J.; Xiong, H.; Zhang, W. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence, Online, 2–9 February 2021; Volume 35*, pp. 11106–11115.
38. Das, A.; Kong, W.; Leach, A.; Mathur, S.; Sen, R.; Yu, R. Long-term forecasting with tide: Time-series dense encoder. *arXiv* **2023**, arXiv:2304.0842.

Disclaimer/Publisher’s Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.