

Transformer Architectures for Multivariate Time Series Forecasting

Deliang Zhang ^a, Bo He*, Yuxin Gao ^b

College of Computer Science and Engineering, Chongqing University of Technology, Chongqing 400054, China

^a814705867@qq.com; *Corresponding author: hebo@cqut.edu.cn; ^b2365720153@qq.com

Abstract: Multivariate time series data constitutes the fundamental basis for decision making across real-world scenarios such as financial trading, intensive care, energy dispatch and the Internet of Things. Traditional statistical methods and early deep learning architectures including recurrent neural networks and convolutional neural networks expose structural bottlenecks when these models capture long-range dynamic dependencies, process high-dimensional cross correlations and accommodate data expansion on a massive scale. The Transformer architecture relies on the self-attention mechanism to break the traditional paradigm of sequence modeling through parallel computing capabilities and flexible global receptive fields. This architecture has emerged as the cutting-edge foundation for the analysis of multivariate time series. However, the standard Transformer model remains restricted by quadratic computational complexity, data starvation effects and the lack of interpretability associated with black box characteristics when this model processes real-world data featuring high-frequency fluctuations, non-stationarity and significant noise interference. This paper systematically synthesizes the key theoretical and methodological breakthroughs in the field of time series forecasting in recent years. This work starts from the underlying feature representation mechanism to deeply deconstruct data representation strategies which include dimensionality reduction decomposition, time-frequency domain conversion and masked pre-training. By integrating specific application scenarios, this paper conducts a fine-grained dynamic classification and performance evaluation of current deep learning models. This evaluation specifically focuses on efficient sparse attention, sequence decomposition, patch-based designs and multimodal hybrid Transformer architectures. Building upon this foundation, this work examines the severe challenges facing the field and prospectively explores future academic trajectories which encompass time series foundation models, physics-informed design, counterfactual reasoning and federated learning. The ultimate goal of this paper is to provide a panoramic reference for the theoretical breakthroughs and engineering deployment of complex time series analysis.

Keywords: Multivariate time series; Deep learning; Transformer; Attention mechanism; Feature extraction; Foundation models.

1. Introduction

Time series analysis occupies the central position of data science research serving as a crucial method for uncovering objective physical laws and socioeconomic evolution. The rapid development of modern industrial environments and digital platforms has generated massive and highly complex multivariate time series (MTS) data. Unlike isolated univariate sequences, MTS data involve multiple interdependent variables where the evolution process frequently intertwines with high-dimensional cross-coupling effects, multi-scale dynamic cycles and unpredictable non-stationary abrupt changes [1]. Spanning from the macroeconomic indicator fluctuations of financial markets and the multi-source physiological signals of critical patients in medical monitoring to the load curves of smart grids affected by both meteorology and scheduling, MTS forecasting goes far beyond simple trend extrapolation and serves as the fundamental support for driving forward-looking decisions in high-risk and highly dynamic scenarios. The precise extraction of potential causal relationships and long-range spatiotemporal dependencies from heterogeneous data streams that contain redundant noise, exhibit missing values and have varying sampling frequencies constitutes a severe challenge for current forecasting tasks.

In the early research stages, time series forecasting relied heavily on statistical models based on strict mathematical assumptions. Methods such as autoregressive integrated

moving average (ARIMA) and exponential smoothing process non-stationarity through differencing and smoothing operators, which have been successfully applied to macroeconomic and demographic forecasting characterized by linear relationships. However, when facing large-scale high-dimensional data generated by Internet of Things sensors, such models are restricted by linear and stationary assumptions, making it difficult to capture intricate nonlinear coupling features. With the leap in computing power and the emergence of big data, the sequence modeling paradigm based on deep learning has gradually replaced traditional statistical models. Recurrent neural networks (RNN) and variants including LSTM and GRU introduce gating mechanisms and hidden state memory units to greatly enhance the capability of the model to retain historical information, which can autonomously extract abstract temporal dynamic features from original sequences. Almost simultaneously, temporal convolutional networks (TCN) utilize dilated causal convolution structures to expand the temporal receptive field in a more efficient parallel computing manner [2]. Nevertheless, when the prediction horizon expands and variable dimensions surge, the serialized processing bottleneck and gradient attenuation problems of the RNN architecture remain difficult to eradicate. Concurrently, the CNN architecture is limited by a fixed-length receptive field and appears inadequate in modeling cross-variable long-range dependencies, where the performance of the model often exhibits a suboptimal state as

the sequence length increases.

The advent of the Transformer architecture has triggered a profound paradigm shift in the field of sequence modeling [3]. This architecture completely abandons the recursive constraints of recurrent and convolutional structures and relies on the Multi-Head Self-Attention mechanism to establish global interconnections within sequences. This mechanism not only breaks the restriction of sequence length on computational parallelism but also endows the model with the ability to dynamically assign weights to different time steps and feature dimensions to capture complex spatiotemporal dependencies. Although the original Transformer has achieved disruptive success in the field of natural language processing, direct transplantation to MTS forecasting exposes defects of poor adaptation. The continuity of time series data fundamentally differs from the discrete symbolic characteristics of natural language, where the quadratic computational complexity () of the original self-attention mechanism causes severe memory and computing power bottlenecks when processing long sensor sequences of tens of thousands of steps. Furthermore, the standard Transformer lacks inductive bias for the periodicity and trend of time series, making it highly susceptible to overfitting when dealing with vertical domain applications with limited data volumes. Additionally, pure self-attention weights cannot provide reliable clinical or financial interpretability evidence for high-risk scenarios [4].

Aiming at the aforementioned inherent architectural defects, the academic community has set off a wave of modifications for the time series adaptability of the Transformer. Researchers have reconstructed network topologies and feature processing logic from multiple dimensions, which has spawned numerous highly pioneering variant models. On the one hand, introducing probabilistic sparse attention algorithms including Informer significantly reduces the complexity boundary to make extreme long-sequence forecasting possible [5]. On the other hand, integrating the classical seasonal-trend decomposition concept with deep networks including Autoformer and FEDformer guides the model to separately process macroscopic evolution and microscopic oscillations in different subspaces of the time domain and frequency domain [6]. Recently, the segmentation strategy based on local patches further effectively alleviates the signal-to-noise ratio imbalance problem in multivariate forecasting by aggregating local semantics and independently processing each channel [7]. These evolutions not only reflect the transition of model structures from pure attention stacking to integrating prior structures but also signify that MTS analysis is marching towards a new era of large model pre-training and zero-shot inference [8].

Based on the diversity and complexity of current technological evolution, this paper aims to construct a systematic evaluation and prospective framework for multivariate time series forecasting. This work transcends the traditional review perspective of singly enumerating model architectures and instead deconstructs the learning logic of deep time series features from the combined dimension of underlying data representation and high-level network abstraction. This paper not only dynamically outlines the evolutionary mapping of deep architectures focusing on Encoder-Decoder, pure Transformer, hybrid networks and generative adversarial paradigms, but also deeply analyzes data augmentation and representation methods including

dimensionality reduction decomposition, time-frequency conversion and contrastive masked pre-training. By mapping algorithmic theories to physical application scenarios such as medical care, energy and transportation, this paper systematically reveals deployment barriers including parallel computing bottlenecks, heterogeneous data fusion and model interpretability. Furthermore, this work attempts to provide structured academic coordinates for next-generation evolutionary paths that encompass physics-informed embedding, counterfactual reasoning mechanisms and time series foundation models.

2. Theoretical Foundations and Challenges of Multivariate Time Series

Real-world multivariate time series (MTS) data represent complex signals formed by the superimposition of multiple potential dynamic mechanisms rather than irregular chaotic values. Establishing the mathematical components of time series and analyzing the corresponding forecasting barriers constitute a prerequisite for building robust architectures before deep neural networks can effectively extract the high-dimensional features of the data.

2.1. Constituent Elements of Time Series

Theoretically, a standard continuous observation sequence can usually be decoupled into three independent and interacting components which include a trend term, a seasonal term and a residual term [9]. The trend component maps the global evolution direction of the data over a long observation period and reflects macroscopic physical laws which include the continuous expansion of economies and the irreversible aging of industrial equipment. Correspondingly, the seasonal component depicts the regular oscillations driven by specific time cycles which include days, weeks and months. This recurring pattern occupies a dominant position in the analysis of tidal traffic flow and climate energy dissipation. However, residual components that cannot be explained by deterministic mechanisms inevitably remain in the sequence after the clear macroscopic trends and periodic patterns are stripped away. These residuals (or white noise) integrate the inherent random fluctuations of the system and environmental background interference. Furthermore, the internal autocorrelation features of the residuals provide indispensable statistical clues for evaluating the goodness of fit of the model and capturing unobserved hidden variables [10]. By decoupling the highly entangled original signals into basic components along the time dimension, researchers can more accurately isolate deterministic knowledge and uncertainty risks to establish clear representation boundaries for subsequent network learning.

2.2. Core Challenges of MTS Forecasting

Although classical sequence decomposition theory provides an intuitive physical picture for understanding the evolution of univariate signals, the complexity of forecasting tasks exhibits a leapfrog growth when the research vision expands to multi-source heterogeneous multivariate time series. MTS forecasting in real environments suffers from the dual challenges of spatial dimension explosion and temporal dynamic drift.

2.2.1. Computation of High-Dimensional Feature Spaces and Overfitting Traps

The variable dimension of MTS surges exponentially with the rapid expansion of the scale of collection nodes and sensor networks. This high-dimensional characteristic is not a simple stacking of data volume but the direct root cause that triggers the depletion of computing resources and representation collapse. Traditional algorithms face the severe curse of dimensionality in high-dimensional subspaces. This challenge makes it difficult to perform effective feature visualization and dimensionality reduction operations. Moreover, network models are highly prone to capturing spurious correlations in the complex variable covariance matrix, which sharply amplifies the risk of model overfitting [11]. Accurately filtering out the core feature dimensions that dominate system evolution while effectively controlling computational complexity constitutes the primary bottleneck of current network architecture optimization. This optimization must avoid the quadratic complexity disaster associated with the standard self-attention mechanism.

2.2.2. Dynamic Nonlinear Dependencies Between Variables

The increase in dimensionality is inevitably accompanied by the complication of deep interactive relationships between variables. In MTS systems, variables in various dimensions do not exist in isolation but are coupled with each other through hidden system dynamic mechanisms. These coupling correlations are extremely challenging because they generally exhibit highly nonlinear features and undergo dynamic evolution over time alongside phase shifts across time steps [12]. For example, in macroeconomic regulation scenarios, the interactions among inflation rate, interest rate and GDP will undergo structural remodeling in different market cycles. Deep models must possess the ability to untangle these dynamic interactive topological networks. Ignoring the synergistic evolution across variables or treating this evolution as a static prior relationship will inevitably lead to the distortion of long-term forecasting paths.

2.2.3. Data Alignment Difficulties and Sparsity Interference

Data collection mechanisms in actual deployment environments depart from idealized mathematical assumptions and further exacerbate the difficulty of information extraction. Sensors across different physical distributions struggle to guarantee absolute clock synchronization. This issue causes MTS data to generally face alignment and sparsity dilemmas which include irregular sampling intervals, asynchronous measurements and large-scale missing values. Different variables exhibit inherent lagging or leading effects when these variables respond to external shocks. Traditional linear interpolation or zero-padding methods often destroy the original local trend and frequency features of the sequence and subsequently introduce structural deviations into the deep representation stage [13]. When large-area data sparsity occurs in key observation windows, the model faces a severe fault of historical information. Furthermore, the statistical inference of the model derived on incomplete manifolds is highly prone to falling into local optima.

2.2.4. Temporal Dynamic Evolution and Non-stationarity

The multi-scale evolution and non-stationary properties of the time dimension itself constitute the final barrier that hinders the generalization ability of MTS forecasting models.

Real-world time series systems rarely maintain strict statistical stationarity. The potential data distribution of these systems will experience severe concept drift and distribution shifts alongside system state switching, strong external interventions or sudden black swan events which include sudden policy changes and extreme meteorological disasters [14]. This multi-dimensional temporal dynamism requires forecasting architectures to possess a multi-resolution receptive field that balances long-term gentle attenuation and short-term high-frequency mutations. Additionally, these architectures must exhibit extremely strong robustness when facing out-of-distribution samples. If the architecture lacks the inductive bias to capture such non-stationary evolution, models that achieve extremely low fitting errors on closed training sets will inevitably face a precipitous decline in performance when these models encounter the dynamic shocks of real deployment environments.

3. Classification of Deep Learning and Transformer Model Architectures

In the field of multivariate time series (MTS) forecasting, deep learning architectures have experienced a paradigm shift from relying on local receptive fields and recursive states to possessing global perception and long-range modeling capabilities. To deeply understand this evolutionary trajectory, this section deconstructs existing forecasting networks into two major camps which include conventional deep learning architectures and the core dominant Transformer architectures, and this section conducts a fine-grained analysis of the internal topological variants of these models.

3.1. Evolution of Conventional Deep Learning Architectures

Before the Transformer established a dominant position, conventional deep learning models occupied an important place in sequence modeling through different topological designs. As an early foundational paradigm, the Encoder-Decoder model and the typical Seq2Seq structure of this model map input sequences into latent abstract representations through long short-term memory (LSTM) networks or gated recurrent units (GRU), after which the decoder autoregressively generates multi-step forecasting results [15]. Although this architecture performs excellently in capturing historical dependencies, highly symmetric network structures (such as the U-Time model adapted from medical image segmentation [16]) often require the input and output spaces to remain highly consistent. This consistency requirement inevitably encounters the bottleneck of information attenuation when the model faces complex multivariate mapping.

Generative adversarial networks (GAN) have opened another path for sequence feature learning through adversarial games. Models represented by TimeGAN ingeniously combine the flexibility of unsupervised learning and the control of supervised learning to synthesize realistic time series data through the dynamic game between generators and discriminators [17]. Furthermore, RCGAN based on a recursive conditional architecture demonstrates unique generative advantages in capturing nonlinear features and heavy-tailed distributions of highly volatile sequences such as financial data [18]. However, the application of GANs in MTS forecasting is often limited by training instability and mode collapse problems, and these networks are more

frequently downgraded and applied to data augmentation and missing value imputation.

To bridge the defects of single architectures, ensemble and cascade models have emerged. These architectures advocate separating the tasks of local and global feature extraction. By fusing the sensitivity of convolutional neural networks (CNN) to local spatial patterns and the temporal memory capacity of recurrent neural networks (RNN), these models build a deep hybrid representation space. Cascade models go one step further by stacking identical network modules or specifically designed residual blocks (such as the cascade mechanism of TreeDRNet) to ensure effective gradient backpropagation while increasing network depth. This approach approximates complex nonlinear spatiotemporal dynamic manifolds in iterations.

3.2. Evolution and Breakthrough of Transformer Architectures

Conventional architectures gradually expose the fatal weaknesses of low parallel efficiency and limited receptive fields when dealing with extremely long sequences and high-dimensional features, whereas the introduction of the Transformer has completely reshaped the underlying logic of MTS modeling. However, the standard Transformer is not impeccable. The core self-attention mechanism of this model requires pairwise inner products for all nodes in the sequence during computation. This quadratic time and space complexity () directly locks the hardware feasibility of long-term forecasting. More crucially, the native Transformer architecture is permutation-invariant, and relying solely on simple positional encoding cannot provide the necessary temporal continuity and periodic inductive bias for MTS data.

Aiming at the shackles of computing power and feature capturing, long-sequence efficient Transformers take the lead in achieving breakthroughs in the sparsification of attention matrices. The Informer model creatively proposes the ProbSparse self-attention mechanism, which evaluates the activity of queries and keys based on Kullback-Leibler divergence and retains only the connections with the highest information entropy, successfully compressing the computational complexity to [5]. This milestone improvement not only eliminates the memory bottleneck of long-sequence modeling but also prompts subsequent research to push the optimization of attention mechanisms to the extreme.

Simultaneously, researchers realized that pure attention sparsification could not untangle the complex dynamic phenomena intertwined in the data, which led to the emergence of decomposition-based Transformers. Autoformer directly embeds the sequence decomposition strategy into network layers to spontaneously strip macroscopic trends and microscopic seasonal terms internally. This model adopts an autocorrelation mechanism to replace traditional dot-product attention, which significantly improves the capture efficiency for periodic patterns [6]. FEDformer pushes the decomposition concept into the frequency domain. By combining seasonal-trend decomposition with Fourier or wavelet transforms, this model completes feature aggregation in the frequency domain subspace. This approach not only filters high-frequency noise but also achieves outstanding long-range forecasting performance with linear complexity [19].

Recently, foundation models based on patches and pre-training represent the latest evolutionary direction of MTS

architectures. PatchTST divides continuous long sequences into non-overlapping or overlapping local patches. This method multiplies the reduction of input sequence length and lowers computational overhead while greatly preserving the integrity of local semantic information [7]. With the penetration of the large model concept, time series foundation models possessing zero-shot generalization capabilities (such as TimeGPT-1) have begun to show prominence [8]. Such models rely on massive cross-domain datasets for pre-training and demonstrate general sequence understanding capabilities that are stripped of specific task constraints.

Although pure Transformer architectures have dominance in global dependency modeling, these models are often less efficient than convolutional networks in capturing fine-grained local features. Therefore, hybrid architectures become an inevitable choice to bridge this gap. By deeply fusing the Transformer with CNNs, GNNs or RNNs, hybrid architectures make up for the shortcomings of pure attention mechanisms in multi-scale representation and local noise filtering while maintaining macroscopic spatiotemporal insights. Examples of such architectures include CT-PatchTST which utilizes channel-time joint perception modules to strengthen new energy forecasting [20] and PatchTST-GRU which combines numerical weather prediction for multi-step wind power forecasting [21].

4. Feature Extraction and Representation Learning Enhancement Methods

Beyond precise network architectures, how to extract the deep representations with the most predictive value from high-dimensional data sources mixed with noise, non-stationary drift and missing values constitutes the core determining the upper limit of MTS models. Modern deep forecasting paradigms have shifted from early end-to-end black box mapping to feature enhancement methodologies that highly conform to the inherent physical and statistical laws of time series.

4.1. Dimension Decomposition and Time-Frequency Conversion

The appearance of complex time series is often an illusion formed by the superimposition of multiple potential dynamic mechanisms. Dimension decomposition serves as a powerful prior method which advocates explicitly deconstructing original high-dimensional tensors into components with clear physical meanings including trends, seasonality and residuals. This process not only eliminates the severe perturbation of non-stationary trends on the gradients of the model but also enables the network to allocate heterogeneous parameter spaces for components of different frequencies. For instance, the LaST model utilizes variational inference to decouple seasonal and trend representations in the latent space, which significantly improves the recognition accuracy of the model for potential laws and the transparency of the output [22].

Compared to difficult pattern matching in the time domain, time-frequency conversion provides an orthogonal perspective for feature extraction. By mapping time-domain signals to the frequency domain, models can intuitively strip transient high-frequency noise and accurately target the core periodic oscillation features that drive sequence evolution. The FreTS model directly constructs multi-layer perceptrons (MLP) in the frequency domain, which not only accelerates

the convergence of cross-channel correlations but also effectively alleviates the over-smoothing problem of time-domain self-attention [23]. Furthermore, the FiLM model combines Fourier analysis and low-rank matrix approximation to achieve highly efficient feature denoising and information reconstruction in the frequency domain [24].

4.2. Self-Supervised Pre-training Strategies

Because time series possess lower information density compared to natural language and the acquisition cost of high-quality annotated data is exorbitant, traditional supervised learning is highly prone to overfitting on small sample sets. Self-supervised pre-training provides a powerful representation learning framework for dealing with data scarcity and domain shifts by mining the structural context of the data itself.

Contrastive learning plays a key role in this framework. The core logic of this method lies in constructing positive and negative sample pairs on spatiotemporal manifolds. By pulling enhanced homologous samples closer and pushing non-homologous samples away in the representation space, this approach forces the encoder to learn invariant features that resist perturbations. The TF-C model innovatively requires representations in the time domain and frequency domain to maintain consistency in a joint space. This model utilizes rich spectral information to strengthen contrastive loss, thereby providing an extremely robust feature foundation for downstream fine-tuning tasks [25].

On the other hand, masked modeling directly confronts the ubiquitous problems of asynchronous sampling and sensor failures in actual MTS data. By artificially applying random masking to input sequences and forcing the network to reconstruct missing information, models are compelled to master dynamic dependencies across time steps and the covariance topology among multiple variables. Relying on masked autoencoder architectures, models such as STEP and Ti-MAE effectively enhance the tolerance to data uncertainty and extreme noise during the reconstruction process, demonstrating outstanding generalization and transfer capabilities [26, 27].

4.3. Patch-Based Segmentation and Channel Independence

Traditional point-wise inputs often destroy the complete semantics of local patterns in time series and cause extreme expansion of sequence length. Patch-based segmentation thoroughly breaks these shackles. By slicing long sequences into a series of patches equipped with local contextual semantics, models acquire denser feature representations at the token level.

A more disruptive approach is the channel independence strategy tightly integrated with the patch mechanism. When facing high-dimensional MTS data, forcing models to compute the cross correlations of all variables within the same attention matrix often leads to severe feature entanglement and redundant interference. The channel independence strategy advocates completely stripping multiple variables, where all channels share the same set of Transformer backbone parameters for independent evolution. This strategy which seemingly cuts off cross-variable interactions actually proves extremely effective in resisting the noise amplification effects brought by useless features. This approach not only greatly compresses the number of parameters of the model but paradoxically reaches the pinnacle of performance in

multivariate forecasting across numerous benchmark tests [7].

5. Core Application Domains and Benchmark Datasets

The theoretical evolution of deep learning and Transformer architectures ultimately requires validation in highly complex real-world scenarios. Physical laws and data generation mechanisms across different fields exhibit profound differences. This characteristic introduces entirely different challenges and focuses when time series forecasting is implemented in practice.

5.1. Core Application Scenarios of Multivariate Time Series

In financial and stock markets, asset price sequences intertwine with extremely high volatility and non-stationarity and are constantly subjected to severe impacts from external noise including macroeconomic indicators and market sentiment [28]. Traditional autoregressive models struggle to capture long-term price trends under such extreme signal-to-noise ratios. Networks introducing residual enhancement and channel spatial attention mechanisms (such as RCSAN) can effectively extract multi-scale features of financial data to lock onto potential evolution patterns in dynamically fluctuating markets [29]. In stark contrast, forecasting tasks in the healthcare domain (such as vital sign monitoring in intensive care units) are often hindered by irregular sampling of data and extremely strong individual physiological differences. In this scenario, zero-shot generalization models (such as TimelyGPT) successfully overcome the bottleneck of missing patient historical records through pre-training on large-scale medical sequences. This approach demonstrates immense potential in processing high-dimensional heterogeneous physiological signals [30].

Energy systems and transportation networks represent typical scenarios driven by spatiotemporal topologies and strong external covariates. Electricity load forecasting and the power generation estimation of renewable energy sources heavily depend on meteorological data variations and long-period seasonal patterns. Urban traffic flow estimation is constrained by complex road network spatial graph topologies and the spreading patterns of dynamic congestion [31]. In these fields, pure time-domain attention appears inadequate, which requires models to possess spatiotemporal joint modeling capabilities to analyze cross-regional synergistic evolution. Furthermore, in the anomaly detection scenarios of the Internet of Things, the core demand shifts to accurately identifying rare events and periodic deviations from massive and multi-source sensor data streams. This characteristic requires models to possess extremely strong anti-noise robustness while maintaining high sensitivity to conventional patterns [13].

5.2. Benchmark Datasets and Diversified Evaluation Systems

The algorithmic validation of the aforementioned complex scenarios highly depends on standardized benchmark datasets that possess broad consensus. The current academic community has constructed data matrices covering multiple domains. The most representative ones include the Electricity Transformer Temperature (ETT) dataset used to capture extreme long-range dependencies, the California highway sensor network dataset (PEMS) used to verify high-

dimensional spatiotemporal graph associations [32], and the Exchange-rate and Weather datasets that encompass complex macroscopic dynamics. Meanwhile, datasets provided by large-scale international forecasting competitions such as M4 and M5 have become touchstones for measuring the generalization ability of models because these datasets cover rich modalities ranging from microscopic commodity sales to macroeconomic indicators [33].

As forecasting architectures become increasingly complex, the evaluation system has also evolved from a single accuracy orientation to multi-dimensional composite indicators. Traditional statistical indicators based on errors remain the cornerstone for measuring the goodness of fit of models. Mean squared error (MSE) and root mean squared error (RMSE) apply quadratic penalties to large-scale forecasting deviations, which possess decisive reference value in scenarios that have zero tolerance for extreme errors including energy grid mutations. Mean absolute error (MAE) provides a more intuitive measure of global deviation. However, absolute errors often generate severe evaluation biases when models face cross-dimensional heterogeneous datasets in finance and healthcare. Symmetric mean absolute percentage error (SMAPE) and MAPE exhibit better fairness in evaluating cross-modal data due to their scale-invariant characteristics [28]. More crucially, given the inherently high computational overhead of the Transformer architecture, modern evaluation systems have mandatorily introduced considerations for the number of model parameters, peak memory consumption and online inference latency. This comprehensive trade-off between accuracy and efficiency is guiding the academic community from the misunderstanding of blindly stacking network depth back to a pragmatic track oriented towards edge deployment and real-time computing [13].

6. Current Technical Challenges

Although deep networks represented by the Transformer have repeatedly broken records in major benchmark tests, these architectures still expose numerous unbridged underlying theoretical and engineering implementation gaps when these models migrate to industrial-level deployment and safety-critical systems.

First, computational scalability and computing power costs constitute the most direct physical ceiling. The quadratic complexity exhibited by the standard self-attention mechanism when facing sequence lengths causes models to rapidly reach the bottlenecks of video memory overflow and computing power exhaustion when processing long-period high-dimensional data with tens of thousands of time steps in energy or climate analysis [28, 13]. Although the academic community has proposed compromise solutions including sparse attention or dimensionality reduction patching, these methods often come at the expense of the integrity of global temporal dependencies. How to achieve attention expression with linear complexity while guaranteeing absolute accuracy remains an unsolved mathematical problem.

Second, data starvation and the overfitting phenomenon under small sample conditions greatly limit the downstream application of models. The massive parameter space of modern deep architectures requires massive data for feeding. However, in data-scarce fields including rare disease forecasting or specific industrial fault detection, forcefully fitting high-capacity models easily leads networks to capture spurious correlations of environmental noise and data

specificity rather than authentic physical dynamic mechanisms. This vulnerability, which performs excellently on training sets but faces precipitous collapse on unknown distributions, severely weakens the cross-scenario generalization ability of models [31, 13].

Furthermore, the lack of interpretability acts as the core Achilles heel that hinders deep time series models from entering high-risk decision-making links. Deeply stacked feedforward layers and intricate multi-head attention weight matrices encapsulate the forecasting process into a highly nonlinear black box. In medical diagnosis and financial risk assessment, decision makers not only need to know future forecasting points but also need to clarify the causal deduction logic behind the conclusions [31]. Existing attention weight visualizations are often filled with noise and cannot be directly equivalent to true feature importance. This characteristic deprives model outputs of sufficient persuasiveness when facing regulatory scrutiny or clinical validation.

Simultaneously, the separation between pure data-driven paradigms and specific domain knowledge exacerbates the blindness of model outputs. The vast majority of current Transformer models treat time series as pure numerical sequences for abstract fitting while disregarding the prior physical laws (such as energy conservation and thermodynamic equations) or macroeconomic structural constraints that govern system operations behind the data. This modeling approach lacking physical consistency not only wastes valuable domain prior knowledge but also easily deviates from common sense when forecasting long-term horizons, which generates absurd conclusions that are completely untenable according to physical laws [28, 13].

Finally, the fragmentation of benchmark tests and the crisis of reproducibility are eroding the academic credibility of this field. Different literatures often adopt highly inconsistent look-back window lengths, forecasting step sizes and data normalization strategies in comparative experiments. Certain studies exhibit severe selection bias in hyperparameter tuning and even obtain artificially high metrics through data leakage. This chaotic evaluation environment lacking unified and fair comparison protocols not only conceals the authentic gains brought by architectural innovations but also makes it difficult for successors to accurately locate the true pros and cons of various models within the existing technology landscape [13].

7. Future Research Directions

Although existing models have achieved substantial progress in multivariate time series forecasting, a broad exploration space remains to construct universal, robust and interpretable time series artificial intelligence. The theoretical breakthroughs and engineering evolution of the field in the future will highly depend on the deep integration of cross-disciplinary mechanisms.

7.1. Time Series Foundation Models and Zero-shot Inference

With the maturity of large-scale pre-training techniques in the field of natural language processing, constructing cross-domain time series foundation models has become an inevitable trend to overcome the challenge of data starvation. Future research focuses will shift towards extracting time series representations with high transferability. This shift aims

to conduct massive self-supervised learning on heterogeneous datasets including meteorology, finance and healthcare through a unified Transformer backbone, thereby endowing models with zero-shot forecasting capabilities under unseen data distributions [13, 31]. This large model paradigm is expected to completely eliminate the computational redundancy of repeated fine-tuning for single tasks.

7.2. Physics-Informed Modeling and Domain Knowledge Embedding

Purely data-driven black box mappings are facing a crisis of reliability, which requires the next generation of forecasting architectures to realize the leap from statistical fitting to physical mechanism constraints. By implicitly embedding partial differential equations, laws of thermodynamics or macroeconomic topological priors into the attention matrix of the Transformer, constructing physics-informed models can force networks to generate forecasting trajectories that conform to realistic common sense. This deep integration of domain knowledge will significantly enhance the extrapolation robustness of models under extreme external shocks [28, 13].

7.3. Spatiotemporal Fusion and Multimodal Collaborative Learning

Dynamic systems in the real world rarely evolve in isolation with a single modality. Breaking the perspective limitations of one-dimensional time series and aligning high-dimensional heterogeneous modalities in a unified attention space constitute a new frontier of multimodal collaborative learning. These modalities encompass textual sentiment including financial news, spatial topology including traffic road network nodes and visual signals including meteorological satellite clouds [31, 13]. Cross-modal Transformers will possess a broader perception ability for system environments, which provides a more three-dimensional and richer decision context for multi-scale forecasting under complex network environments.

7.4. Counterfactual Reasoning and Privacy Protection Mechanisms

To compensate for the inherent deficiency of pure correlation analysis in logical rigor, deeply nesting causal inference frameworks with the Transformer to generate actionable counterfactual explanations represents a key path to breaking the trust barriers of high-risk applications [31]. Furthermore, in data silo scenarios with strong regulations including healthcare and financial transactions, relying on federated learning to realize cross-institutional decentralized Transformer collaborative training acts as the core institutional guarantee for promoting the large-scale deployment of deep time series models. This approach completes the parameter aggregation of global forecasting knowledge while ensuring the absolute physical isolation of underlying sensitive data [31, 13].

8. Conclusion

The evolutionary history of multivariate time series forecasting essentially represents a technological game against the curse of dimensionality, dynamic drift and signal-to-noise ratio imbalances. Relying on the globally interconnected self-attention mechanism and highly expansive representation capacity, the Transformer

architecture has completely subverted the old sequence modeling paradigm dominated by recurrence and local convolution to establish a core dominant position in long-range complex time series analysis. However, secondary crises including the explosive growth of computational costs, restricted interpretability and disconnection from physical meanings all serve as warnings that the simple stacking of architectures has inevitably reached the limit of marginal returns.

The path to future breakthroughs is destined to seek an exquisite balance between model capacity and computational efficiency alongside the balance between purely data-driven approaches and physical structure constraints. From the sparse decomposition improvement of self-attention matrices to the macroscopic reconstruction of multimodal spatiotemporal fusion and zero-shot large models, multivariate time series forecasting is accelerating towards the stage of artificial general intelligence equipped with strong causal explanatory power and cross-domain adaptability. This evolution will not only profoundly reshape the decision centers of core industrial lifelines including finance, healthcare and energy, but also provide an unprecedented cognitive engine for humans to gain insight into and control the evolutionary laws of complex dynamic systems.

Acknowledgements

This work is funded by the humanities and social sciences research project of the ministry of education (No.24YJA870003).

References

- [1] Box, G. E. P., Jenkins, G. M., Reinsel, G. C., & Ljung, G. M. (2015). *Time series analysis: Forecasting and control*. John Wiley & Sons.
- [2] Bai, S., Kolter, J. Z., & Koltun, V. (2018). An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv*, arXiv:1803.01271. <https://doi.org/10.48550/arXiv.1803.01271>
- [3] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. In *Advances in Neural Information Processing Systems 30 (NeurIPS)* (pp. 5998-6008).
- [4] Lim, B., Arik, S. Ö., Loeff, N., & Pfister, T. (2021). Temporal fusion transformers for interpretable multi-horizon time series forecasting. *International Journal of Forecasting*, 37(4), 1748-1764. <https://doi.org/10.1016/j.ijforecast.2021.03.012>
- [5] Zhou, H., Zhang, S., Peng, J., Zhang, S., Li, J., Xiong, H., & Zhang, W. (2021). Informer: Beyond efficient transformer for long sequence time-series forecasting. In *Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 35, No. 12, pp. 11106-11115)*. AAAI Press. <https://doi.org/10.1609/aaai.v35i12.17325>
- [6] Wu, H., Xu, J., Wang, J., & Long, M. (2021). Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. In *Advances in Neural Information Processing Systems 34 (NeurIPS)* (pp. 22419-22430).
- [7] Nie, Y., Nguyen, N. H., Sinthong, P., & Kalagnanam, J. (2022). A time series is worth 64 words: Long-term forecasting with transformers. *arXiv*, arXiv:2211.14730. <https://doi.org/10.48550/arXiv.2211.14730>
- [8] Garza, A., & Mergenthaler-Canseco, M. (2023). TimeGPT-1. *arXiv*, arXiv:2310.03589. <https://doi.org/10.48550/arXiv.2310.03589>

- [9] Shumway, R. H., & Stoffer, D. S. (2000). Time series analysis and its applications. Springer.
- [10] Blázquez-García, A., Conde, A., Mori, U., & Lozano, J. A. (2021). A review on outlier/anomaly detection in time series data. *ACM Computing Surveys*, 54(3), Article 56. <https://doi.org/10.1145/3444690>
- [11] Liu, P., Lin, J., & Zhang, C. (2024). Heterogeneous multivariate functional time series modeling: A state space approach. *IEEE Transactions on Knowledge and Data Engineering*, 36(12), 8421-8433. <https://doi.org/10.1109/TKDE.2024.3402765>
- [12] Wang, Z., Fan, J., Wu, H., Sun, D., & Wu, J. (2024). Representing multiview time-series graph structures for multivariate long-term time-series forecasting. *IEEE Transactions on Artificial Intelligence*, 5(6), 2651-2662. <https://doi.org/10.1109/TAI.2023.3332564>
- [13] Saravana, M. K., Roopa, M. S., Arunalatha, J. S., & Venugopal, K. R. (2026). Transformers for multivariate time series forecasting: Comprehensive analysis, challenges, research opportunities, and future prospects. *IEEE Access*, 14, 11424-11457.
- [14] Xu, H. (2025). AFMDS: Attention-based forecasting model against distribution shift for long-term non-stationary time series prediction. *IEEE Access*, 13, 155797-155807.
- [15] Sutskever, I., Vinyals, O., & Le, Q. V. (2014). Sequence to sequence learning with neural networks. *Advances in Neural Information Processing Systems*, 27, 3104-3112.
- [16] Perslev, M., Jensen, M., Darkner, S., Jennum, P. J., & Igel, C. (2019). U-Time: A fully convolutional network for time series segmentation applied to sleep staging. In *Advances in Neural Information Processing Systems 32 (NeurIPS)*.
- [17] Yoon, J., Jarrett, D., & van der Schaar, M. (2019). Time-series generative adversarial networks. In *Advances in Neural Information Processing Systems 32 (NeurIPS)*.
- [18] Esteban, C., Hyland, S. L., & Rätsch, G. (2017). Real-valued (medical) time series generation with recurrent conditional GANs. *arXiv*, [arXiv:1706.02633](https://doi.org/10.48550/arXiv.1706.02633). <https://doi.org/10.48550/arXiv.1706.02633>
- [19] Zhou, T., Ma, Z., Wen, Q., Wang, X., Sun, L., & Jin, R. (2022). FEDformer: Frequency enhanced decomposed transformer for long-term series forecasting. In *Proceedings of the 39th International Conference on Machine Learning (ICML)* (pp. 27268-27286). PMLR.
- [20] Lu, K., Huo, M., Li, Y., Zhu, Q., & Chen, Z. (2025). CT-PatchTST: Channel-time patch time-series transformer for long-term renewable energy forecasting. *arXiv*, [arXiv:2501.08620](https://doi.org/10.48550/arXiv.2501.08620). <https://doi.org/10.48550/arXiv.2501.08620>
- [21] Xu, S., Wang, Y., Xu, X., Shi, G., Huang, H., & Zheng, Y. (2025). A PatchTST-GRU based heterogeneous seq2seq model with numerical weather prediction refinement for multi-step wind power forecasting. *Scientific Reports*, 15(1), Article 16547. <https://doi.org/10.1038/s41598-025-98765-4>
- [22] Wang, Z., Xu, X., Zhang, W., Trajcevski, G., Zhong, T., & Zhou, F. (2022). Learning latent seasonal-trend representations for time series forecasting. In *Advances in Neural Information Processing Systems 35 (NeurIPS)* (pp. 38775-38787).
- [23] Yi, K., Zhang, Q., Fan, W., Wang, S., Wang, P., He, H., Lian, D., An, N., Cao, L., & Niu, Z. (2023). Frequency-domain MLPs are more effective learners in time series forecasting. *arXiv*, [arXiv:2311.06184](https://doi.org/10.48550/arXiv.2311.06184). <https://doi.org/10.48550/arXiv.2311.06184>
- [24] Zhou, T., Ma, Z., Wen, Q., Sun, L., Yao, T., Yin, W., & Jin, R. (2022). FiLM: Frequency improved Legendre memory model for long-term time series forecasting. In *Advances in Neural Information Processing Systems 35 (NeurIPS)* (pp. 12677-12690).
- [25] Zhang, W., Yang, L., Geng, S., & Hong, S. (2022). Self-supervised time series representation learning via cross reconstruction transformer. *arXiv*, [arXiv:2205.09928](https://doi.org/10.48550/arXiv.2205.09928). <https://doi.org/10.48550/arXiv.2205.09928>
- [26] Shao, Z., Zhang, Z., Wang, F., & Xu, Y. (2022). Pre-training enhanced spatial-temporal graph neural network for multivariate time series forecasting. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining (KDD)* (pp. 1567-1577). ACM. <https://doi.org/10.1145/3534678.3539396>
- [27] Li, Z., Rao, Z., Pan, L., Wang, P., & Xu, Z. (2023). Ti-MAE: Self-supervised masked time series autoencoders. *arXiv*, [arXiv:2301.08871](https://doi.org/10.48550/arXiv.2301.08871). <https://doi.org/10.48550/arXiv.2301.08871>
- [28] Zhao, J., Chu, F., Xie, L., Che, Y., Wu, Y., & Burke, A. F. (2026). A survey of transformer networks for time series forecasting. *Computer Science Review*, 60, Article 100883. <https://doi.org/10.1016/j.cosrev.2026.100883>
- [29] Sun, W., Liu, Z., Yuan, C., Zhou, X., Pei, Y., & Wei, C. (2025). RCSAN: Residual enhanced channel spatial attention network for stock price forecasting. *Scientific Reports*, 15(1), Article 21800. <https://doi.org/10.1038/s41598-025-87654-3>
- [30] Song, Z., Lu, Q., Xu, H., Zhu, H., Buckeridge, D., & Li, Y. (2024). TimelyGPT: Extrapolatable transformer pre-training for long-term time-series forecasting in healthcare. In *Proceedings of the 15th ACM International Conference on Bioinformatics, Computational Biology and Health Informatics* (pp. 1-10). ACM. <https://doi.org/10.1145/3699767.3699791>
- [31] Kong, X., Chen, Z., Liu, W., Ning, K., Zhang, L., Marier, S. M., Liu, Y., Chen, Y., & Xia, F. (2025). Deep learning for time series forecasting: A survey. *International Journal of Machine Learning and Cybernetics*, 16, 5079-5112. <https://doi.org/10.1007/s13042-025-02345-6>
- [32] Chen, C., Petty, K., Skabardonis, A., Varaiya, P., & Jia, Z. (2001). Freeway performance measurement system: Mining loop detector data. *Transportation Research Record*, 1748(1), 96-102. <https://doi.org/10.3141/1748-12>
- [33] Makridakis, S., Spiliotis, E., & Assimakopoulos, V. (2018). The M4 competition: Results, findings, conclusion and way forward. *International Journal of Forecasting*, 34(4), 802-808. <https://doi.org/10.1016/j.ijforecast.2018.06.001>