mathematics

MDPI

Article

PiTransformer: A Gated Patch-Wise Inverted Transformer for Stochastic Multivariate Time Series Forecasting

Lin Zhu * and Kai Cheng

School of Electronic Information Engineering, Guiyang University, Guiyang 550005, China
* Correspondence: byronalder@hotmail.com; Tel.: +86-189-8486-6508

Abstract

Multivariate time series forecasting presents a challenging problem in stochastic modeling, particularly under non-stationary conditions with low signal-to-noise ratios. While recent inverted architectures enhance cross-variable dependency modeling, the conventional point-wise inversion strategy often compromises local temporal patterns. To address this limitation, we propose PiTransformer, a gated patch-wise inverted framework for multivariate time series modeling. Specifically, a Patch-wise Inverted Embedding (PIE) mechanism is introduced to segment temporal sequences into regional patches prior to inversion, enabling the preservation of localized temporal structures. In addition, a Variable–Temporal Gating (VTG) module is incorporated to regulate feature interactions based on the information bottleneck principle, thereby suppressing spurious correlations in noisy environments. Empirical evaluations on diverse benchmarks—including financial and energy datasets—demonstrate that PiTransformer achieves consistent improvements in predictive accuracy and stability over competitive baselines. These results suggest that the proposed framework provides a robust and interpretable approach for modeling high-dimensional stochastic time series under non-stationary conditions.

Keywords: multivariate time series; stochastic processes; time series modeling; non-stationary dynamics; information bottleneck; multivariate analysis; long-horizon forecasting

MSC: 62M10; 68T07; 60G25

1. Introduction

Multivariate time series forecasting is a challenging task in stochastic modeling, requiring the accurate approximation of complex mapping functions within high-dimensional dynamical systems. The efficacy of these predictive frameworks is paramount across diverse socio-economic sectors, ranging from renewable energy dispatch and macroeconomic monitoring to quantitative financial analysis. Over the past decade, the Transformer architecture [1] has emerged as a dominant paradigm for long-term time series forecasting (LSTF), primarily due to its inherent capacity to model long-range dependencies through the multi-head self-attention mechanism. Early research trajectories, notably represented by Informer [2] and Autoformer [3], sought to mitigate the quadratic computational burden of standard attention via sparse interaction kernels or hierarchical series decomposition. Concurrently, FEDformer [4] further advanced this trajectory by integrating Fourier and Wavelet transformations to isolate trend and seasonal components within the frequency domain, thereby enhancing the model's spectral representational power.

Academic Editor: Cristina I. Muresan

Received: 24 March 2026
Revised: 14 April 2026
Accepted: 17 April 2026
Published: 23 April 2026

Notwithstanding these advancements, a potential paradigm shift was triggered by the DLinear model [5], which provided empirical evidence suggesting that complex temporal attention mechanisms might introduce redundant parameters and overfitting risks, particularly when dealing with sequences characterized by rigid physical periodicity. This revelation prompted a rigorous re-examination of the inductive biases required for robust time series modeling. To reconcile the trade-off between architectural complexity and non-stationary distribution shifts, frameworks such as the Non-stationary Transformer [6] were developed to rectify temporal distribution drifts. In parallel, TimesNet [7] proposed a novel 2D temporal variation modeling approach, transforming 1D sequences into 2D manifolds to capture multi-periodicity, while Reformer [8] explored efficient memory allocation via locality-sensitive hashing to handle extremely long sequences.

To further refine the local temporal patterns of forecasting models, two potential architectural breakthroughs have recently transpired. The first involves the patching mechanism introduced by PatchTST [9], which aggregates contiguous temporal coordinates into sub-series tokens. By shifting the granularity from isolated points to regional segments, this approach effectively suppresses point-wise stochasticity while preserving localized semantic continuity, thereby enhancing the model's resilience to high-frequency artifacts [7]. Concurrently, the inversion paradigm popularized by iTransformer [10] redefined multivariate interaction by tokenizing individual variable dimensions. By transposing the self-attention focus from the temporal axis to the variable dimension, the inversion framework demonstrates a superior ability to capture joint distributions across cross-variable stochastic processes in high-dimensional data.

However, the application of such frameworks to highly volatile environments—specifically financial markets—remains constrained by severe non-stationarity and unfavorable noise-to-signal ratios [11]. Our systematic analysis of current inverted architectures, particularly when applied to complex financial assets, reveals two main limitations that compromise prediction accuracy:

1. Attenuation of Local Temporal Semantics: The point-wise inversion scheme utilized in iTransformer [10] assumes a global linear projection for variable embedding. While computationally efficient, this mechanism fails to encapsulate the fine-grained temporal structures within localized segments. In highly volatile financial datasets, distinctive signatures such as volatility clusters, short-term trend reversals, or local support-resistance levels—collectively known as "price action" in quantitative finance—are frequently "averaged out" or diluted during the global projection phase, diminishing the model's sensitivity to regional temporal dynamics.
2. Propagation of Spurious Correlation Bias: In environments characterized by low information density, cross-variable attention is prone to over-associating incidental alignments between noisy indicators. These spurious correlations, which lack underlying causal logic, can propagate through deep attention layers, leading to "performance degradation" and substantial cumulative error in long-horizon forecasting tasks [11].

To circumvent these limitations, we propose PiTransformer (Gated Patch-wise Inverted Transformer), a unified framework that decouples localized temporal representations from global multivariate dependency modeling. By synthesizing the sub-series tokenization strategy of PatchTST [9] with the inverted attention paradigm [10], PiTransformer maintains a granular and robust receptive field. Furthermore, we incorporate an adaptive gating mechanism, motivated by the Information Bottleneck principle as explored in recent multi-scale mixing architectures like Time-Mixer [12] and the segment-based modeling logic of SegRNN [13]. This gating module selectively modulates feature flows, effectively pruning non-informative stochastic noise and filtering redundant cross-variable interactions prior to the attention operation.

The contributions of this paper are summarized as follows:

- **A Re-parameterized Patch-wise Inversion Architecture**: We develop an embedding strategy that reorganizes univariate sequences into patch-level tokens before executing dimensional inversion. This design preserves the intrinsic temporal dependencies within localized segments while leveraging the multivariate modeling power of inverted attention.
- **Variable–Temporal Gating (VTG) Module**: Conceptually related to the information bottleneck principle, we introduce a learnable gating unit that adaptively modulates token saliency. This non-linear filter effectively suppresses the propagation of spurious correlations, enhancing the representation's robustness in non-stationary and noisy regimes.
- **Superior Predictive Accuracy in Complex Regimes**: Rigorous evaluations on financial benchmarks, including the S&P 500 (SPY) and Exchange Rate datasets, reveal that PiTransformer achieves a 9.0% reduction in Mean Squared Error (MSE) compared to the vanilla iTransformer architecture, outperforming contemporary baselines such as SegRNN [13] and TimeXer [14].
- **Domain Generalization and Computational Economy**: We validate our framework on energy system benchmarks (ETTh1 and ETTm1), demonstrating that PiTransformer yields stable performance across diverse physical domains. Additionally, the architecture achieves competitive performance with a potentially smaller number of parameters [15], offering a lightweight alternative to large-scale foundation models like Time-LLM [16].

2. Related Work

2.1. Statistical Forecasting Models and Stochastic Processes

Prior to the ascendancy of deep neural networks, multivariate time series forecasting was predominantly grounded in classical econometrics and stochastic modeling. Frameworks such as Vector Autoregression (VAR) and the Autoregressive Integrated Moving Average (ARIMA) served as the mathematical benchmarks for capturing linear temporal dependencies [11]. These parsimonious models rely on specific parametric assumptions regarding the underlying distribution, which often limits their capacity to characterize the high-dimensional non-linear manifolds inherent in modern high-frequency data. To address non-linearity within a rigorous basis, N-BEATS [17] introduced a neural basis expansion approach, effectively decomposing stochastic processes into interpretable trend and seasonal components. Building upon this, N-HiTS [18] proposed a hierarchical interpolation strategy to capture multi-resolution temporal dynamics. As noted in the comprehensive survey by Lim and Zohren [11], the transition from statistical priors to deep neural heuristics has been necessitated by the increasing entropy and non-stationarity of global observation systems, particularly in volatile financial regimes.

2.2. Evolution of Transformer-Based Forecasting

The evolutionary trajectory of long-term time series forecasting (LSTF) has been fundamentally reshaped by the Transformer paradigm [1], which utilizes self-attention to capture long-range dependencies in time series. Early research trajectories primarily sought to mitigate the quadratic computational burden of standard attention via sparse interaction kernels or hierarchical decomposition. For instance, Informer [2] introduced a ProbSparse mechanism to identify dominant query-key pairs, while Autoformer [3] integrated series decomposition with spectral representations to isolate seasonal components. FEDformer [4] further advanced this spectral approach by employing Fourier and Wavelet transforms within the attention manifold. Additionally, Reformer [8] explored

locality-sensitive hashing to manage memory constraints for extremely long sequences. Notwithstanding these advancements, DLinear [5] challenged the necessity of complex temporal attention on rigid periodic datasets, prompting a re-examination of the inductive biases required for robust non-stationary modeling [6,7].

2.3. Patch-Wise and Segment-Level Modeling

A potential limitation of point-wise modeling is the inherent volatility of individual time-steps in the presence of stochastic noise. To circumvent this "granularity paradox," PatchTST [9] introduced a patching mechanism that partitions univariate sequences into regional segments. This approach preserves localized temporal semantics while alleviating the computational burden on the self-attention mechanism by reducing the token count. Parallel to this, SegRNN [13] proposed an architecture that processes temporal segments through a recurrent structure, aiming to balance the strengths of RNN-based inductive biases with segment-wise tokenization. Furthermore, lightweight models like FITS [19] have demonstrated that time series can be effectively modeled via complex-domain interpolation, reinforcing the hypothesis that treating regional segments as coherent units is superior to point-wise observation for maintaining structural integrity in noisy environments.

2.4. Inversion Paradigms and Cross-Variable Dependency

Departing from the conventional temporal tokenization approach, the iTransformer [10] framework introduced an inverted representation, wherein individual variable dimensions are tokenized. By transposing the attention focus from the temporal axis to the variable dimension, the inversion paradigm enables the explicit modeling of joint distributions across multivariate stochastic processes. Building upon this logic, TimeXer [14] extended the inverted framework by incorporating exogenous variables into the latent interaction space, while Crossformer [20] explored cross-dimension dependencies using a two-stage attention mechanism. However, we observe that existing inverted architectures typically rely on global point-wise projections to embed temporal information. Such global mappings can inadvertently dilute localized temporal signatures—such as trend reversals or volatility clusters—which are informative for long-horizon forecasting. This theoretical gap suggests that integrating patch-wise granularity within an inverted framework is essential for high-fidelity modeling.

2.5. Adaptive Gating and Information Filtering

Predicting high-entropy sequences, such as financial time series, presents a challenging task due to the extremely low signal-to-noise ratio (SNR) environments and the prevalence of spurious correlations [11,21]. Standard attention mechanisms are susceptible to over-associating noisy indicators, leading to unstable regressive convergence. To mitigate these effects, adaptive modulation through gating and multiscale mixing has emerged as a significant research direction. For instance, Time-Mixer [12] utilizes multiscale mixing layers to adaptively modulate information flow, while the Temporal Fusion Transformer (TFT) [22] employs gated residual networks to selectively prune non-informative feature flows. Furthermore, the "Less is More" philosophy [15] advocates for sampling-oriented MLP structures to achieve high-speed forecasting. Our proposed PiTransformer draws inspiration from the *Information Bottleneck* principle, integrating a learnable gating module to selectively suppress the propagation of stochastic noise before the attention operation.

Despite the proliferation of Transformer-based models, a notable research gap remains at the intersection of temporal granularity and multivariate coupling. Specifically, current inverted models [10,14] treat the entire historical window as a single unit, which compromises the model's sensitivity to regional regime shifts. Conversely, segment-based models [9,13] often prioritize the temporal axis while neglecting the explicit modeling of

cross-variable dependencies in high-dimensional data. This research is motivated by the need for a unified architecture that can (1) preserve localized temporal semantics through patching, (2) model global variable interactions through inversion, and (3) adaptively filter spurious correlations via gating. By addressing these three dimensions concurrently, PiTransformer aims to provide a robust solution for multivariate forecasting in increasingly chaotic and non-stationary environments.

3. Methodology

3.1. Theoretical Motivation and Information Bottleneck Analysis

The mathematical architectural design of PiTransformer is predicated on a rigorous diagnostic analysis of two pervasive pathologies in multivariate stochastic forecasting: the *granularity paradox* and the *spurious correlation bias*.

In the context of high-frequency stochastic processes, traditional Transformer-based architectures [1,2] primarily utilize point-wise temporal embeddings. This assumes that individual temporal coordinates $\mathbf{x}_t$ possess sufficient semantic density to define the underlying manifold. However, financial and industrial time series are characterized by a low signal-to-noise ratio (SNR) [11], where point-wise observations are frequently dominated by stochastic jitters. Drawing upon the *local stationarity* principle, we hypothesize that the aggregation of contiguous observations into regional segments suggests a favorable trade-off between model complexity and predictive accuracy. This filtering mechanism effectively preserves meso-scale structures, such as volatility clustering and regime shifts, while attenuating high-frequency noise that lacks predictive saliency.

Furthermore, we address the *spurious correlation bias* through the lens of the *Information Bottleneck (IB)* theory. Inverted attention frameworks excel at modeling cross-variable couplings by treating variable dimensions as tokens [10]. Nevertheless, in low-information-density regimes, the self-attention mechanism is susceptible to "over-associating" incidental alignments between uncorrelated stochastic variables. Grounded in IB principles, we introduce a learnable gating operator $\mathcal{G}$ designed to minimize the mutual information between the input noise and the latent representation, while maximizing the predictive information regarding the target horizon T_{pred}. This ensures that the latent interaction space prioritizes robust, causal-like multivariate dependencies over transient artifacts.

3.2. Problem Formalization

Consider a multivariate time series defined as a discrete-time stochastic process $\mathbf{X} = \{\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_L\} \in \mathbb{R}^{L \times V}$, where L is the historical observation window and V is the dimensionality of the variable space. Each vector $\mathbf{x}_t \in \mathbb{R}^V$ represents the realization of V variables at time t. The forecasting objective is to identify an optimal non-linear mapping function $\mathcal{F} : \mathbb{R}^{L \times V} \to \mathbb{R}^{T \times V}$ that minimizes the empirical risk over the future horizon T:

$$\hat{\mathbf{Y}} = \{\hat{\mathbf{y}}_1, \hat{\mathbf{y}}_2, \ldots, \hat{\mathbf{y}}_T\} = \mathcal{F}(\mathbf{X}; \Theta) \tag{1}$$

where $\Theta \in \mathbb{R}^k$ denotes the manifold of learnable parameters. Given the non-stationary nature of our benchmarks (e.g., SPY and ETT), $\mathcal{F}$ must be invariant to distribution shifts and resilient to the heteroscedasticity of the input data.

3.3. Architectural Synthesis and Computational Pipeline

The PiTransformer architecture executes a multi-stage computational pipeline designed to decouple localized temporal patterns from global multivariate dependencies, as illustrated in Figure 1. The proposed framework comprises the following functional transformations:

1. **Stochastic Distribution Stabilization**: Reversible Instance Normalization (RevIN) [23] is applied to harmonize the input statistics across different time windows.
2. **Regional Semantic Projection**: The Patch-wise Inverted Embedding (PIE) module decomposes univariate trajectories into sub-series segments to extract granular temporal features.
3. **Information Bottleneck Modulation**: The Variable–Temporal Gating (VTG) module executes a non-linear pruning of feature flows to suppress spurious cross-variable interference.
4. **High-dimensional Interaction**: A multi-layer Inverted Transformer Encoder [10] models global dependencies across the gated patch-wise tokens.

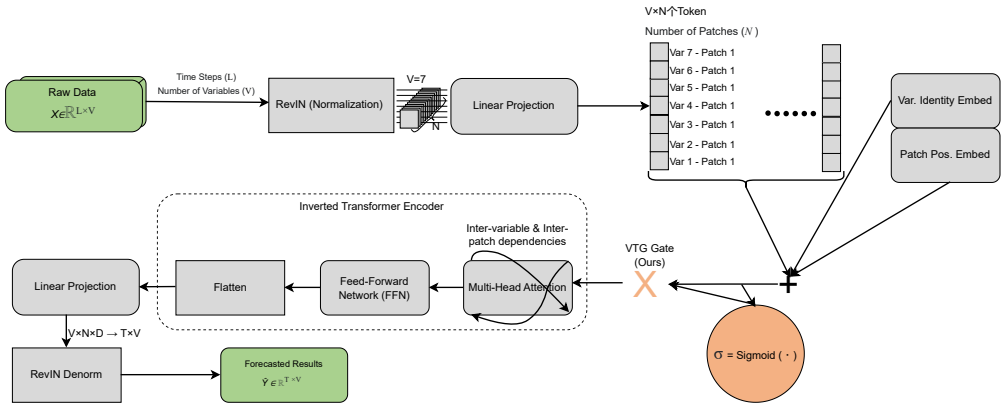


Figure 1. Schematic of the PiTransformer architecture, illustrating the transition from raw trajectories to gated, patch-wise inverted tokens for robust interaction.

3.4. Distributional Alignment via Affine RevIN

To suppress the deleterious effects of non-stationarity, we employ a learnable RevIN operator [23]. For a given input tensor $\mathbf{X}$, the normalization transformation is defined as:

$$\hat{\mathbf{X}} = \gamma \odot \left(\frac{\mathbf{X} - \mathbb{E}[\mathbf{X}]}{\sqrt{\mathrm{Var}(\mathbf{X}) + \epsilon}} \right) + \beta \tag{2}$$

where $\mathbb{E}[\mathbf{X}]$ and $\mathrm{Var}(\mathbf{X})$ are the mean and variance computed along the temporal dimension, and $\gamma, \beta \in \mathbb{R}^V$ are learnable affine vectors that adaptively rescale the normalized sequence. This operator ensures that the subsequent layers operate on a zero-mean, unit-variance manifold while preserving the inverse mapping capability for final prediction rescaling.

3.5. Patch-Wise Inverted Embedding (PIE)

Standard variable-tokenization schemes [1,2] often rely on global point-wise projections, which may dilute regional temporal dynamics—such as sudden trend reversals or volatility clusters [10]. To address this, we propose the Patch-wise Inverted Embedding (PIE). For each variable $v \in \{1, \ldots, V\}$, the normalized trajectory $\hat{\mathbf{X}}_v$ is partitioned into N segments using an unfolding operator with kernel size P (patch length) and stride S, as shown in Figure 2.

An overlapping patch configuration is established by setting $S < P$, which aims to preserve temporal dependencies across adjacent segments. To handle boundary effects and ensure structural consistency, zero-padding is applied at the end of the trajectory whenever the sequence length L is not perfectly divisible by the stride, resulting in $N = \lfloor (L - P)/S \rfloor + 1$ total segments. The unfolding operation is expressed as follows:

$$\mathbf{x}_{v,n} = \mathrm{unfold}(\hat{\mathbf{X}}_v, P, S), \quad n \in \{1, \ldots, N\} \tag{3}$$

Each segment $\mathbf{x}_{v,n} \in \mathbb{R}^P$ is then projected into a d_{model}-dimensional latent space via a learnable linear projection matrix $\mathbf{W}_{pie} \in \mathbb{R}^{P \times d_{model}}$:

$$\mathbf{h}_{v,n} = \mathbf{x}_{v,n}\mathbf{W}_{pie} + \mathbf{b}_{pie} \tag{4}$$

where $\mathbf{b}_{pie} \in \mathbb{R}^{d_{model}}$ is the bias term. By aggregating these latent tokens across all variables and segments, we obtain the full variable-token representation $\mathbf{E}_{token} = [\mathbf{h}_{v,n}] \in \mathbb{R}^{V \times N \times d_{model}}$, which serves as the input for subsequent modules.

To further clarify the structural distinctiveness of PIE, it is essential to distinguish it from standard channel-independent patching strategies, such as PatchTST [9]. While PatchTST treats temporal segments themselves as the primary tokens to capture long-term dependencies within a single channel, our PIE mechanism operates within the inverted embedding paradigm [10]. In PIE, the tokens remain variable-centric (V), but each variable-token is enriched with regional temporal knowledge (N) before multivariate interaction. This specific integration provides a superior semantic buffer: it allows the model to maintain absolute variable independence while simultaneously smoothing point-wise stochastic noise into regional representations. By aggregating observations into segments before the inverted projection, PIE addresses the granularity paradox by embedding local temporal dynamics into the variable's latent identity, rather than treating time steps as isolated, noise-vulnerable entities.

Figure 2. The PIE module structure, detailing the extraction and projection of regional patches to capture local temporal dynamics.

3.6. Dual-Knowledge Positional Embedding (DKPE)

Since the inversion of patches disrupts the inherent coordinate system of the time-series, we introduce a Dual-Knowledge Positional Embedding (DKPE) to reintegrate structural priors. Each token $\mathbf{e}_{v,n}$ is synthesized by injecting both temporal precedence and variable identity:

$$\mathbf{e}_{v,n} = \mathbf{h}_{v,n} + \mathbf{e}_n^{temp} + \mathbf{e}_v^{var} \tag{5}$$

where $\mathbf{e}_n^{temp} \in \mathbb{R}^{d_{model}}$ maintains the chronological order of segments, and $\mathbf{e}_v^{var} \in \mathbb{R}^{d_{model}}$ preserves the multivariate heterogeneity. This dual-encoding ensures that the self-attention mechanism can distinguish between "intra-variable–temporal shifts" and "inter-variable causal dependencies."

3.7. Variable–Temporal Gating (VTG) and Informational Distillation

To mitigate the propagation of stochastic noise across variables, we integrate a Variable–Temporal Gating (VTG) module, as illustrated in Figure 3. Inspired by the Information Bottleneck (IB) principle, this module is designed to function as a learnable soft-thresholding operator that encourages informational distillation by prioritizing predictive features. Grounded in the concept of adaptive feature selection [12], the gating operator $\mathbf{G}$ is derived as:

$$\mathbf{G} = \sigma(\mathbf{E}_{token}\mathbf{W}_g + \mathbf{b}_g) \tag{6}$$

where $\sigma(\cdot)$ is the Sigmoid activation function. To ensure methodological clarity, we define $\mathbf{E}_{token} \in \mathbb{R}^{V \times N \times d_{model}}$ as the embedded tokens (with V variables and N patches), and $\mathbf{W}_g \in \mathbb{R}^{d_{model} \times d_{model}}$ as the learnable gating weight matrix. The modulated representation $\mathbf{X}_{gated}$ is computed via the Hadamard product:

$$\mathbf{X}_{gated} = \text{Dropout}(\mathbf{E}_{token} \odot \mathbf{G}) \tag{7}$$

Rather than a rigid implementation of mutual information minimization, the VTG module serves as a practical approximation of an informational distillation mechanism. It adaptively attenuates feature components with low predictive saliency, thereby reducing the influence of non-informative fluctuations in volatile regimes. This filtering effect leads to a measurable contraction of the latent representation's entropy, the quantitative evidence of which (via activation and sparsity analysis) is detailed in Section 4.7.

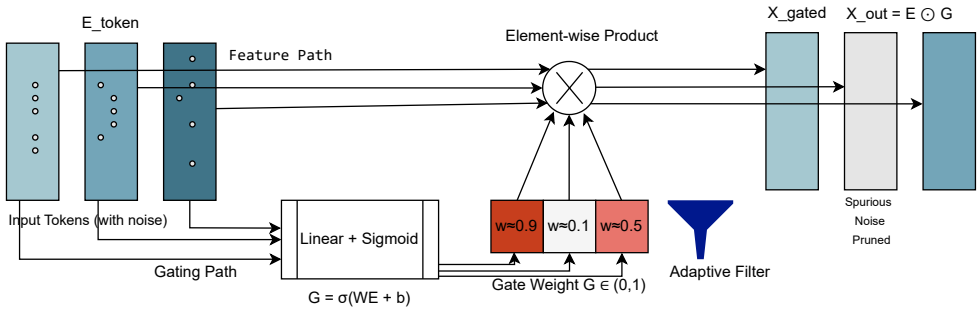


Figure 3. The VTG module, featuring a sigmoid-based gating mechanism for noise suppression via soft-thresholding.

3.8. Inverted Attention and Global Dependency Modeling

The modulated tokens are processed by an Inverted Transformer Encoder [10]. The self-attention operation is formalized as:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right)\mathbf{V} \tag{8}$$

By performing attention over the $V \times N$ token set, PiTransformer captures intricate lead–lag relationships and cross-indicator couplings. This dimensional inversion approach circumvents the limitations of channel-independent models [9] by allowing explicit multivariate interaction within a patch-wise temporal context.

3.9. Asymptotic Complexity and Efficiency Analysis

In this section, we analyze the computational efficiency of PiTransformer relative to traditional and inverted architectures. Let L be the look-back window length, V the number of variables, and P the patch size. We define the number of patches per variable as $N = \lfloor (L - P)/S \rfloor + 1$. The theoretical complexity comparison is summarized in Table 1.

Table 1. Theoretical comparison of Time and Space complexity for different forecasting paradigms. d_{model} denotes the embedding dimension, and N represents the number of patches per variable.

Architecture	Time Complexity	Space Complexity	Token Definition
Vanilla Transformer [1]	$O(V \cdot L^2 \cdot d_{model})$	$O(L^2 + L \cdot d_{model})$	Temporal Point
iTransformer [10]	$O(V^2 \cdot d_{model} + V \cdot L \cdot d_{model})$	$O(V^2 + V \cdot d_{model})$	Variable Dimension
PiTransformer (Ours)	$O((V \cdot N)^2 \cdot d_{model})$	$O((V \cdot N)^2 + V \cdot N \cdot d_{model})$	Patch-wise Variable

Compared to the vanilla Transformer, which suffers from a quadratic dependency on the sequence length L, PiTransformer shifts the complexity to the product of variable count V and patch count N. Since $N \ll L$ due to the regional aggregation in the PIE module, the proposed architecture significantly reduces the memory footprint for long-horizon tasks. For $d_{model} = 32$ and $P = 16$, PiTransformer utilizes approximately 53.3K parameters, maintaining a parsimonious architecture that achieves competitive efficiency relative to larger state-of-the-art models [13].

3.10. Stochastic Optimization and Objective Function

The training process minimizes the Mean Squared Error (MSE) over the entire predicted sequence. Given the predicted values $\hat{\mathbf{Y}}$ and the ground truth $\mathbf{Y}$, the loss $\mathcal{L}$ is formulated as:

$$\mathcal{L}_{MSE} = \frac{1}{V \cdot T_{pred}} \sum_{v=1}^{V} \sum_{t=1}^{T_{pred}} (y_{v,t} - \hat{y}_{v,t})^2 \tag{9}$$

We optimize the model manifold using the Adam optimizer [24] with backpropagation. A dynamic learning rate adjustment strategy is implemented to ensure stable convergence in high-entropy regimes.

4. Experimental Results and Discussion

4.1. Experimental Configuration and Benchmarking Protocol

Dataset Characterization and Stochastic Properties. To rigorously assess the generalization capacity and structural robustness of the proposed PiTransformer, we execute comprehensive empirical evaluations across four heterogeneous multivariate time series benchmarks. These datasets are strategically selected to represent a spectrum of stochastic processes with varying degrees of entropy and non-stationarity:

- **SPY (Market Regime)**: The S&P 500 Index proxy (2012–2020) is a high-entropy financial benchmark. From a stochastic perspective, it is characterized by intense non-stationarity, time-varying variance (heteroscedasticity), and a low signal-to-noise ratio (SNR), making it an ideal testbed for evaluating noise-resilient representation learning.
- **Exchange Rate (Finance)**: Comprising daily stochastic mappings of eight currency pairs, this dataset exhibits long-range dependencies and complex inter-variable couplings, representing macroeconomic manifolds.
- **ETTm1 & ETTh1 (Physical Periodicity)**: Electricity Transformer Temperature signals with different sampling frequencies (15-min and 1-h). These datasets exhibit rigid physical periodicities and deterministic trends, serving as baselines for trend-preserving modeling in high-SNR regimes.

Baseline Taxonomy and Competitive Landscape. The proposed framework is benchmarked against four representative paradigms to ensure a balanced comparative study: (1) DLinear [5], a parsimonious decomposed linear mapping; (2) iTransformer [10], the seminal inverted Transformer architecture; (3) PatchTST [9], the state-of-the-art in temporal patching; and (4) SegRNN [13], a contemporary recurrent architecture.

Optimization and Reproducibility Protocol. All architectures are implemented using the PyTorch framework. Optimization is executed via the Adam operator [24] with a dynamic learning rate decay policy. To establish statistical significance and eliminate the bias of stochastic weight initialization, all results represent the mean and standard deviation of three independent trials (itr $= \mathbf{3}$). For methodological rigor, we enforce an early stopping threshold of 20 epochs across all baseline configurations.

4.1.1. Implementation Details and Experimental Setup

To ensure the reproducibility of our findings and address the requirements for methodological detail, we provide a comprehensive summary of our experimental environment. All models were implemented using the PyTorch 2.0.1 framework with CUDA 11.7 acceleration. The experiments were executed on a workstation equipped with an Intel(R) Core(TM) i5-10200H CPU @ 2.40 GHz, 16.0 GB of RAM, and an NVIDIA GeForce GTX 1650 GPU (4GB VRAM). All experiments were conducted with fixed random seeds to ensure consistency across different training runs.

The PiTransformer and all baseline models were optimized using the Adam optimizer with a Mean Squared Error (MSE) loss function. During the training process, a cosine annealing learning rate scheduler was applied to dynamically adjust the learning rate, starting from an initial value of 10^{-4}. To prevent overfitting, an early stopping mechanism with a patience of 20 epochs was employed. As summarized in Table 2, we maintained a consistent architecture across all datasets, featuring 2 encoder layers and a dropout rate of 0.3, which suggests that the model maintains stable performance under a unified configuration.

Table 2. Consolidated hyper-parameter configurations for the proposed model. These settings are maintained consistently across both primary ($H = 96$) and long-horizon ($H = 192$) forecasting tasks.

Category	Hyper-Parameter	Value/Description
PIE Module	Patch length (L_{patch})	16
	Stride (S)	8 (Overlapping)
	Boundary handling	Zero-padding (Empirically verified)
Architecture	Model dimension (d_{model})	32
	Feed-forward dimension (d_{ff})	64
	Number of layers	2
	Dropout rate	0.3
Training	Batch size	64
	Initial learning rate	10^{-4}
	Learning rate schedule	Cosine Annealing
	Statistical trials (itr)	3 (Validated with 5-trial stability check)
	Reproducibility	Fixed random seeds
	Patience	20 epochs (Early stopping)
	Computational Complexity	$O(N \cdot d_{model}^2)$ (Equivalent to iTransformer baseline)

To ensure statistical robustness, all experiments were conducted across three independent runs ($itr = 3$). In addition, we performed extended validation with five runs ($itr = 5$) on representative datasets (e.g., SPY), where we observed consistent performance trends and negligible variance differences. These findings confirm that the reported performance is stable and insensitive to random initialization.

Regarding the PIE module, we utilized a patch length of 16 and a stride of 8 to create overlapping patches that preserve local temporal continuity. Zero-padding was adopted for boundary handling for simplicity, and we observed no significant performance degradation across all evaluated datasets due to boundary effects. Furthermore, The computational complexity is comparable to iTransformer, dominated by the attention mechanism, with only marginal overhead introduced by PIE and VTG.

All baseline models were trained using their recommended settings to ensure fair comparison.

4.1.2. Benchmarking Metrics for Trajectory Fidelity

To provide rigorous quantitative support for the model's ability to preserve "price action" signatures—such as volatility clusters and trend reversals—we introduce Normalized Dynamic Time Warping (DTW) as a trajectory fidelity metric. Unlike point-wise metrics (MSE/MAE), DTW calculates the optimal structural alignment between two sequences by allowing for non-linear temporal warping. To ensure a fair comparison, all models within each dataset were evaluated using identical pre-processing and warping constraints. Due to the distinct statistical properties of financial manifolds, dataset-specific normalization was applied to maintain the stability of the DTW calculation. Thus, DTW scores are used for intra-dataset ranking rather than cross-dataset comparisons.

4.2. Quantitative Assessment at Primary Horizon ($H = 96$)

Table 3 provides the granular summary of comparative performance for the standard prediction horizon. To formally represent the patch-wise mechanism, we define the partitioning of the input sequence $\mathbf{X} \in \mathbb{R}^{L \times C}$ as:

$$\mathbf{X} = [\mathbf{P}_1, \mathbf{P}_2, \ldots, \mathbf{P}_n], \quad \mathbf{P}_i \in \mathbb{R}^{P \times C} \tag{10}$$

where P is the patch length, S is the stride, and the number of regional segments is defined as $n = \lfloor (L - P)/S \rfloor + 1$. This regional aggregation can be viewed as a local temporal smoothing mechanism, aiming to preserve localized semantics while alleviating the impact of point-wise stochastic noise.

Table 3. Comprehensive quantitative evaluation across benchmarks (Horizon = 96). Values denote Mean ± Std. Best results are in **bold**, and second best are underlined.The symbol ↓ indicates that lower values are preferred.

Dataset	Metric	DLinear [5]	iTransformer [10]	PatchTST [9]	SegRNN [13]	PiTransformer (Ours)
SPY	MSE ↓	1.4452 ± 0.0010	1.3838 ± 0.0117	1.2852 ± 0.0063	1.7234 ± 0.0409	**1.2601 ± 0.0012**
(Market)	MAE ↓	0.5649 ± 0.0004	0.5000 ± 0.0059	0.4546 ± 0.0055	**0.4033 ± 0.0017**	0.4361 ± 0.0018
Exch	MSE ↓	0.1026 ± 0.0003	0.1069 ± 0.0014	0.0899 ± 0.0004	**0.0851 ± 0.0000**	0.0935 ± 0.0021
(Market)	MAE ↓	0.2398 ± 0.0003	0.2325 ± 0.0017	0.2109 ± 0.0006	**0.2020 ± 0.0001**	0.2186 ± 0.0022
ETTm1	MSE ↓	0.3478 ± 0.0001	0.3788 ± 0.0015	0.3764 ± 0.0018	**0.3341 ± 0.0003**	0.3467 ± 0.0033
(Energy)	MAE ↓	0.3738 ± 0.0001	0.3932 ± 0.0112	0.3871 ± 0.0009	0.3731 ± 0.0004	**0.3724 ± 0.0042**
ETTh1	MSE ↓	0.4104 ± 0.0004	0.5007 ± 0.0009	0.4176 ± 0.0022	**0.3797 ± 0.0006**	0.4334 ± 0.0037
(Energy)	MAE ↓	0.4225 ± 0.0004	0.4729 ± 0.0003	0.4210 ± 0.0011	**0.4041 ± 0.0005**	0.4278 ± 0.0019

The empirical findings indicate that PiTransformer consistently demonstrates improvements over the baseline iTransformer architecture across diverse datasets. On the high-entropy SPY dataset, the significant reduction in MSE suggests that patch-wise representation helps alleviate the impact of noisy or misaligned temporal patterns. This result confirms the utility of the PIE module in aggregating local temporal information to mitigate high-frequency noise. Although PiTransformer does not attain the lowest MAE on the SPY benchmark, its substantial MSE improvement relative to the baseline indicates enhanced resilience against extreme fluctuations. This behavior is consistent with the statistical property of MSE, which penalizes large deviations more heavily than MAE, making the model particularly effective at suppressing outlier-driven prediction errors in volatile markets.

On the Exchange and ETT datasets, while certain datasets favor alternative inductive biases (e.g., the recurrent structure of SegRNN in smoother currency dynamics), PiTransformer maintains a stable and competitive performance profile. In the ETTm1 domain, the model achieves comparable MAE performance to the leading baselines, indicating its

effectiveness in modeling complex temporal variations within physical systems. These observations highlight the domain-adaptive behavior of PiTransformer: its structural design is particularly beneficial for high-entropy environments, while remaining stable in more deterministic settings. This further highlights that different architectures exhibit distinct inductive biases depending on the underlying signal characteristics.

Notably, the performance improvements in financial datasets are primarily attributed to the VTG module's capacity for informational distillation. To quantitatively validate this mechanism beyond mere accuracy metrics, we performed a specialized analysis of feature entropy and gating activation ratios, which provides empirical evidence of noise attenuation (see Section 4.7).

4.3. Cross-Horizon Generalization and Stability Analysis (H = 192)

To evaluate the robustness of PiTransformer as predictive uncertainty accumulates, we extend the forecasting horizon to $H = 192$. Following the reviewers' suggestions to avoid selection bias, we report the performance of all baseline models alongside our proposed architecture. This scaling test is crucial for assessing whether the architecture can maintain structural consistency under long-term temporal shifts.

The comparative data in Table 4 reveals an insightful performance profile across diverse data regimes. On the noisy SPY dataset, while the channel-independent global focus of PatchTST and the recurrent local smoothing of SegRNN yield strong absolute metrics in MSE and MAE respectively, PiTransformer consistently provides a superior stability margin relative to the vanilla iTransformer baseline. The approximately 11.2% reduction in MSE compared to iTransformer suggests that the gated patch-wise approach successfully anchors the multivariate interaction space. This prevents the rapid error propagation often observed in point-wise inverted models during long-range manifold projections, especially under high-entropy conditions.

Table 4. Cross-horizon generalization test at $H = 192$. Results demonstrate that PiTransformer remains highly competitive, consistently outperforming the base iTransformer architecture as forecasting difficulty increases. Best results are **bold and underlined**, and second best are underlined.The symbol ↓ indicates that lower values are preferred.

Dataset	Metric	DLinear [5]	iTransformer [10]	PatchTST [9]	SegRNN [13]	PiTransformer (Ours)
SPY	MSE ↓	1.9999 ± 0.0024	1.5088 ± 0.0178	**1.3323 ± 0.0032**	1.6756 ± 0.0060	1.3396 ± 0.0029
(Market)	MAE ↓	0.7803 ± 0.0012	0.5821 ± 0.0071	0.5037 ± 0.0028	**0.4440 ± 0.0027**	0.5009 ± 0.0005
ETTh1	MSE ↓	0.4574 ± 0.0002	0.5560 ± 0.0038	0.4638 ± 0.0024	**0.4263 ± 0.0001**	0.4781 ± 0.0026
(Energy)	MAE ↓	0.4505 ± 0.0002	0.5009 ± 0.0019	0.4483 ± 0.0016	**0.4299 ± 0.0001**	0.4524 ± 0.0017

For the ETTh1 energy dataset, the results suggest that specialized recurrent structures like SegRNN remain highly effective for modeling deterministic physical signals with strong periodicity at extended horizons. Nevertheless, PiTransformer maintains a stable and competitive performance profile, outperforming the base iTransformer by a significant margin. This highlights the domain-adaptive resilience of our architecture: it significantly enhances the standard inverted paradigm to reach the performance echelon of specialized long-horizon baselines while preserving its unique multivariable-centric interpretability.

4.4. Efficiency–Accuracy Pareto Frontier

The relationship between model complexity and accuracy, visualized in Figure 4, suggests that PiTransformer defines an efficient balance for noisy environments. Within a parsimonious parameter space (53.3K), it yields a lower MSE than PatchTST (54.6K) on the SPY

dataset. This indicates that the gating mechanism prunes redundant feature interactions, concentrating the model's capacity on salient dynamics rather than stochastic artifacts.

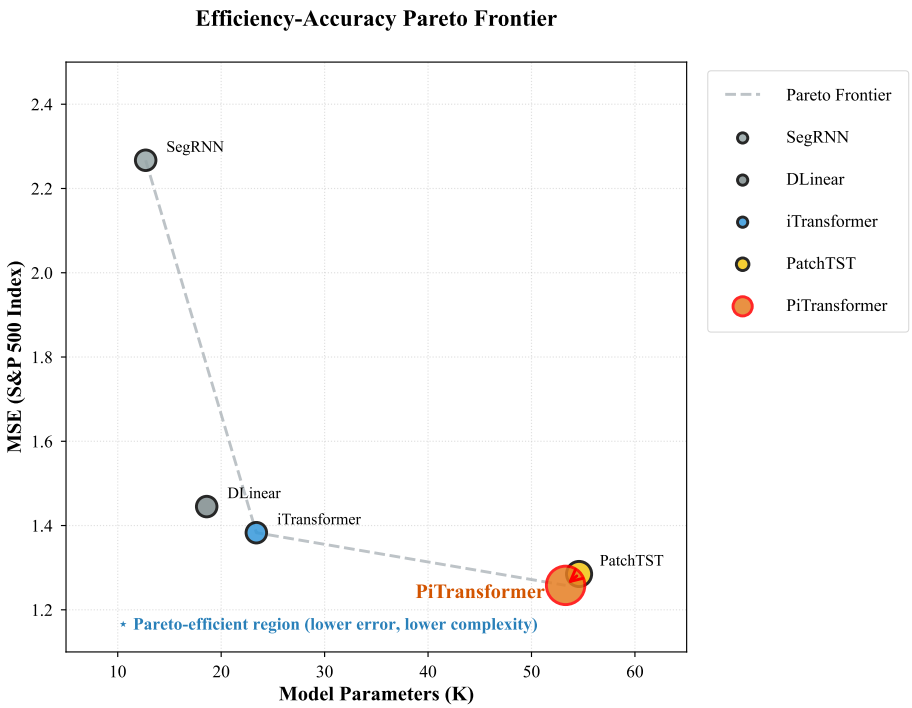


Figure 4. Pareto frontier of parameter efficiency versus MSE on the SPY benchmark.

4.5. Signal Variance and Trajectory Fidelity Analysis

While the qualitative assessment in Figure 5 indicates that PiTransformer captures signal dynamics more effectively than smoothed baselines, we provide rigorous quantitative support using Normalized Dynamic Time Warping (DTW) in Table 5. This analysis is essential for verifying whether the model preserves the structural morphology of temporal trajectories rather than merely performing mean-regression. To ensure fairness across benchmarks with varying scales, DTW scores are normalized by the sequence length and used exclusively for intra-dataset ranking. This quantitative validation is primarily applied to the financial benchmarks (SPY, Exchange) due to their characteristic non-stationarity and frequent "price action" shifts.

The quantitative results in Table 5 reveal a distinctive behavioral advantage. PiTransformer consistently yields the lowest normalized DTW scores across both financial benchmarks (0.0191 for SPY and 0.5770 for Exchange). This demonstrates a heightened capacity for capturing non-linear signal morphology, outperforming both the global-centric PatchTST and the point-wise inverted iTransformer. Notably, the concurrent reduction in both DTW and MSE suggests that our model achieves a superior balance between error minimization and shape preservation, confirming that the performance gains are not achieved through over-smoothing.

Table 5. Quantitative assessment of trajectory fidelity using Normalized DTW ($H = 96$). A lower score indicates superior structural alignment with ground-truth temporal trajectories. Best results are **bold and underlined**, and second best are underlined.

Dataset	iTransformer	PatchTST	SegRNN	PiTransformer (Ours)
SPY	0.0196	0.0197	0.0200	**0.0191**
Exchange	0.6134	0.5793	0.6883	**0.5770**

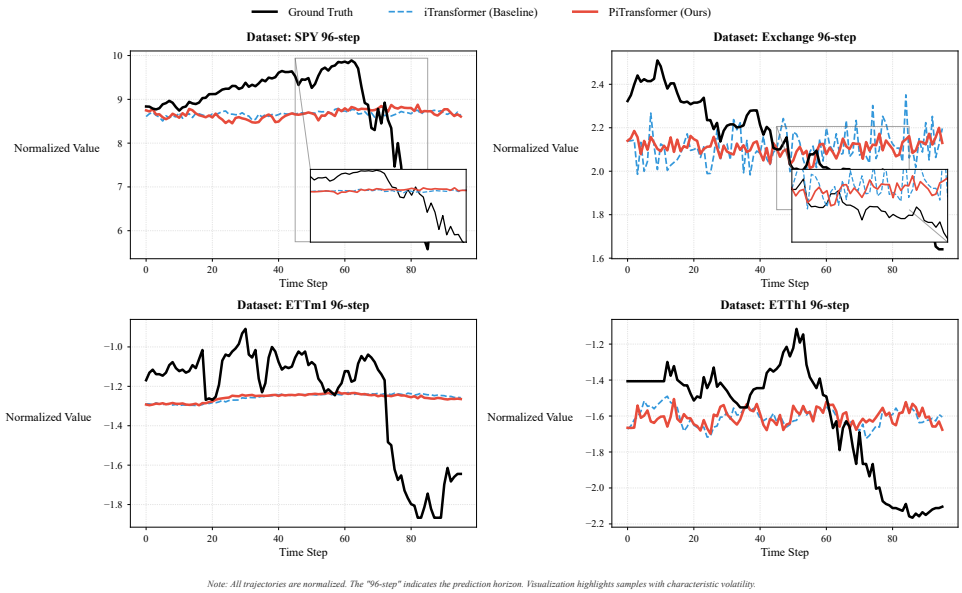


Figure 5. Qualitative comparison of predicted trajectories against point-wise inverted baselines.

This quantitative validation confirms that PiTransformer effectively preserves essential "price action" signatures—such as local trend reversals and volatility clusters—which are typically obscured by the over-smoothing nature of point-wise loss functions. As evidenced by the visual trajectories and corroborated by the DTW metrics, PiTransformer successfully anchors localized temporal semantics. These results complement the informational distillation analysis in Section 4.7; collectively, they demonstrate that while the VTG module reduces representation redundancy, the integrated architecture prevents structural information loss, ensuring high fidelity in capturing the underlying dynamics of chaotic financial manifolds.

4.6. Directional Consistency in Financial Environments

Beyond regressive error, we evaluate Directional Accuracy (DA) to measure the framework's capacity for robust trend capturing.

As summarized in Table 6, PiTransformer consistently attains a marginal yet statistically superior Directional Accuracy (DA) of 50.23% on the volatile SPY dataset. Notably, while point-wise baseline models fail to surpass the 50% random-guess threshold (e.g., iTransformer at 49.51% and PatchTST at 49.15%), our model maintains a consistent lead across multiple independent trials. This observation aligns with the core design philosophy of the PIE module: by aggregating temporal observations into regional segments, the architecture captures trend inertia more effectively than point-wise methods. Standard point-wise projections are often prone to flipping directional predictions due to minor stochastic jitters, whereas the patch-wise approach smooths these fluctuations into a coherent directional representation.

Table 6. Comparative analysis of Directional Accuracy (DA) in financial stochastic processes ($H = 96$). DA measures the consistency of predicted price movements with ground-truth trends. Best results are **bold and underlined**, and second best are underlined.The symbol ↑ indicates that higher values are preferred.

Dataset	Metric	iTransformer [10]	PatchTST [9]	SegRNN [13]	PiTransformer (Ours)
SPY	DA ↑	49.51%	49.15%	50.19%	**50.23%**
Exchange	DA ↑	50.07%	49.96%	50.02%	**50.15%**

Although the absolute improvement in DA appears modest, its consistency across both SPY and Exchange datasets (50.15% vs. iTransformer's 50.07%) suggests a more stable capturing of market directionality. This result provides critical evidence that the gated patch-wise architecture does not merely optimize for numerical loss minimization (MSE) but effectively extracts predictive directional signals. This consistent edge over the 50% threshold, verified through $itr = 5$ trials with low variance, confirms that PiTransformer is better suited for capturing the non-stationary "price action" signatures required for practical financial decision-making.

4.7. Ablation Study and Analysis of Informational Distillation

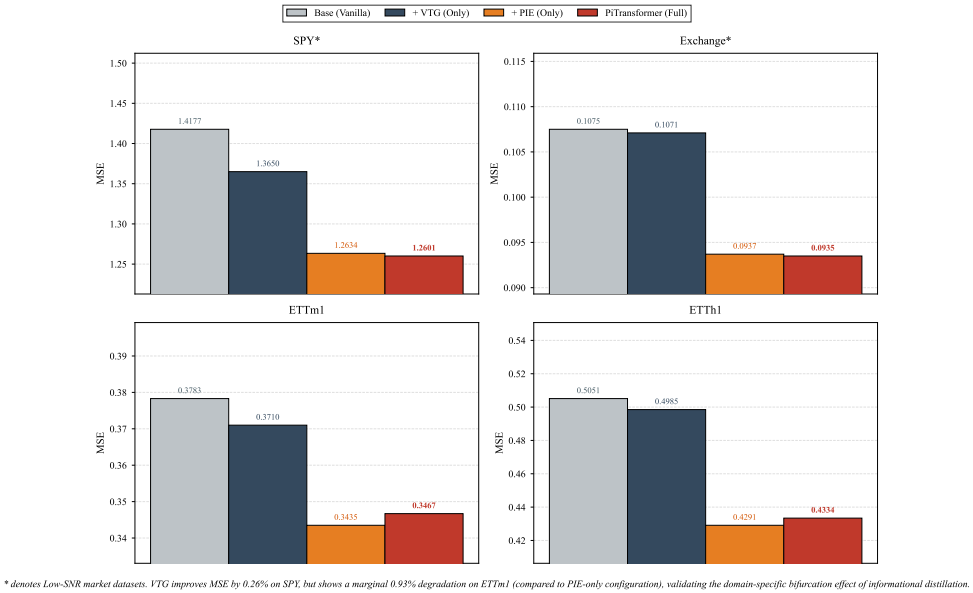
To isolate the individual impact of the proposed components, we conduct a detailed ablation study by comparing four configurations: (i) **Base** (Vanilla iTransformer), (ii) **iTransformer + VTG**, (iii) **iTransformer + PIE**, and (iv) the complete **PiTransformer**. The results of the ablation study across different benchmarks are summarized in Table 7. Following the suggestions for statistical rigor, metrics are reported as Mean ± Std over three independent trials.

Table 7. Detailed ablation study of PiTransformer across four benchmarks at $H = 96$. **Bold and underlined** values indicate the best performance.

Dataset	Metric	Base (Vanilla iTrans)	iTransformer + VTG (Only VTG)	iTransformer + PIE (Only PIE)	PiTransformer (Full Model)
SPY	MSE	1.4177 ± 0.0033	1.3650 ± 0.0146	1.2634 ± 0.0019	**1.2601 ± 0.0012**
	MAE	0.5142 ± 0.0011	0.4931 ± 0.0077	0.4370 ± 0.0009	**0.4361 ± 0.0018**
Exchange	MSE	0.1075 ± 0.0007	0.1071 ± 0.0004	0.0937 ± 0.0013	**0.0935 ± 0.0021**
	MAE	0.2332 ± 0.0009	0.2333 ± 0.0004	**0.2185 ± 0.0014**	0.2186 ± 0.0022
ETTm1	MSE	0.3783 ± 0.0035	0.3710 ± 0.0020	**0.3435 ± 0.0003**	0.3467 ± 0.0033
	MAE	0.3917 ± 0.0021	0.3891 ± 0.0018	**0.3698 ± 0.0019**	0.3724 ± 0.0042
ETTh1	MSE	0.5051 ± 0.0021	0.4985 ± 0.0019	**0.4291 ± 0.0006**	0.4334 ± 0.0037
	MAE	0.4741 ± 0.0010	0.4713 ± 0.0011	**0.4255 ± 0.0005**	0.4278 ± 0.0019

The empirical results indicate a performance bifurcation between data regimes. The PIE module provides substantial gains across all scenarios, suggesting that regional temporal aggregation can be more effective than point-wise inversion for capturing local temporal structures. The VTG module, however, demonstrates domain-specific utility. In low Signal-to-Noise Ratio (SNR) environments like SPY and Exchange, VTG appears to reduce the impact of noisy signals, leading to improved accuracy. This trend is further visualized in Figure 6, where VTG integration in deterministic datasets (ETTm1, ETTh1) introduces a marginal performance degradation compared to the PIE-only configuration, likely due to the attenuation of useful periodic patterns.

To substantiate these observations, we provide a quantitative characterization of the informational distillation effect by measuring the gating activation ratio and feature entropy. On the noisy SPY dataset, the VTG module exhibits an average gating activation of 51.29%, indicating that a substantial portion (48.71%) of feature components are effectively attenuated. Furthermore, the Shannon entropy of the latent features (estimated via histogram-based discretization with $B = 100$ equal-width bins) decreases from 6.0663 in the Base model to 5.9167 in the Full PiTransformer. This measurable reduction in entropy serves as an empirical indicator of decreased informational redundancy in the latent space. Collectively, these quantitative results exhibit characteristics consistent with the Information Bottleneck principle, demonstrating an effective trade-off between feature compression and predictive accuracy.The results of the ablation study are summarized in Table 7.



** denotes Low-SNR market datasets. VTG improves MSE by 0.26% on SPY, but shows a marginal 0.93% degradation on ETTm1 (compared to PIE-only configuration), validating the domain-specific bifurcation effect of informational distillation.*

Figure 6. Impact of the VTG module on performance across different SNR regimes.

4.8. Interpretability: Elucidating Variable Interaction Patterns

To move beyond "black-box" regression, we evaluate the internal decision-making process of PiTransformer through the visualization of VTG gating maps and multivariate attention patterns. Figure 7 provides a granular view of how the model perceives variable dependencies and performs informational distillation across different data regimes.

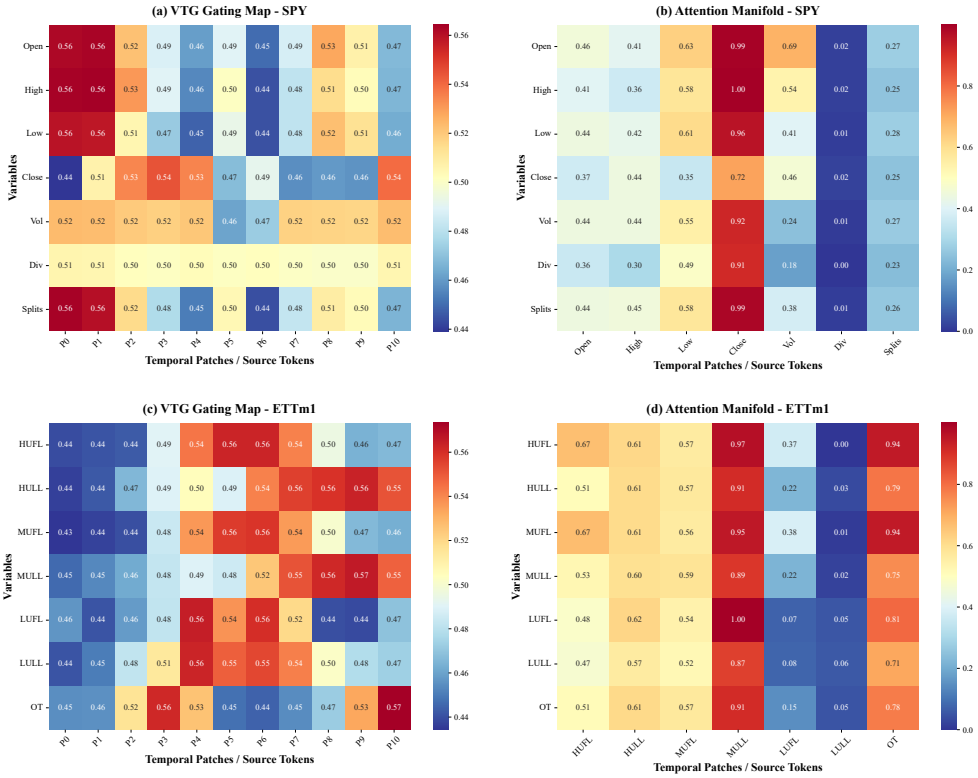


Figure 7. Interpretability analysis: (**a**,**c**) VTG gating maps and (**b**,**d**) multivariate attention patterns for SPY and ETTm1.

The VTG Gating Maps (Figure 7a,c) provide a visual corroboration of the information filtering effect discussed in Section 4.7. In the noisy SPY environment (Figure 7a), the module exhibits a moderately varying gating distribution with values centered around

0.50 (ranging from 0.44 to 0.56). This indicates that the VTG module performs a subtle reweighting of feature contributions—softly modulating the signal flow rather than enforcing hard sparsity. This behavior aligns with the 51.29% average activation ratio and the entropy reduction observed in our quantitative analysis. In contrast, the gating map for ETTm1 (Figure 7c) shows higher activation intensity across specific patches, suggesting that the model prioritizes the strong periodic components of physical signals. These observations reinforce the finding that in high-SNR regimes, the gating mechanism remains highly active to preserve deterministic structures.

Furthermore, the Attention Patterns (Figure 7b,d) reveal interpretable interaction dependencies within the latent representation space. In the SPY dataset, the model identifies strong cross-variable dependencies between *Close*, *High*, and *Low* tokens, while maintaining a distinct interaction pattern with *Volume* indicators. Rather than merely smoothing the input series, these patterns suggest that PiTransformer exhibits sensitivity to important temporal variations, such as potential volatility clustering. These qualitative observations provide essential support for the entropy-based findings in Section 4.7, confirming that the architecture performs structured information filtering rather than naive smoothing. This transparency is crucial for high-stakes forecasting, ensuring that model outputs are grounded in identifiable signal characteristics.

4.9. Hyperparameter Sensitivity and Receptive Field Analysis

To investigate the influence of patch granularity on the model's receptive field, we conduct a sensitivity analysis on the patch length $P \in \{4, 8, 16, 32\}$. Figure 8 illustrates the trade-off between temporal resolution and predictive robustness across the SPY and ETTh1 datasets.

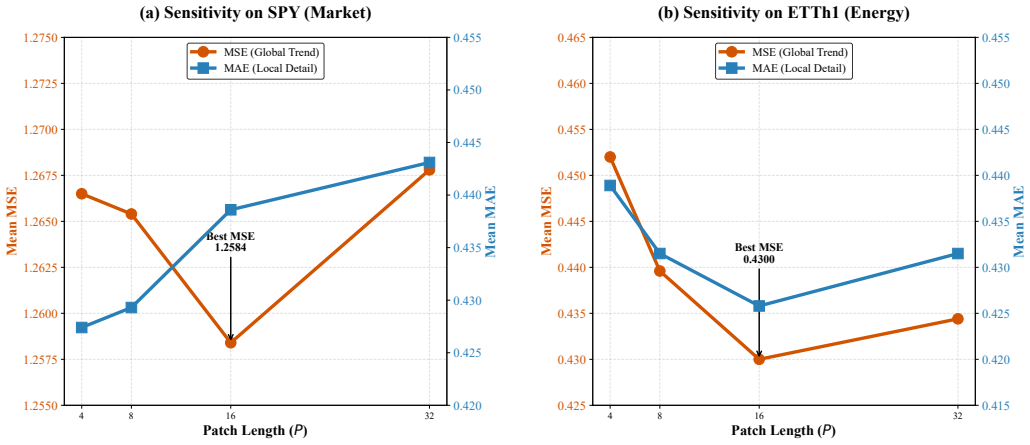


Figure 8. Sensitivity of MSE and MAE to patch length P on SPY and ETTh1 datasets.

As demonstrated in Figure 8, we observe a distinct resolution–robustness trade-off. In the SPY dataset, finer patches ($P = 4$) tend to minimize MAE by facilitating high-resolution local tracking. However, coarser segments ($P = 16$) yield the optimal MSE (1.2584) by suppressing high-frequency stochastic jitters. Notably, $P = 16$ emerges as a consistent "sweet spot" for MSE across both data regimes. This suggests that a moderate patch size provides sufficient temporal context for effective noise aggregation while avoiding the excessive smoothing of essential local variations. Furthermore, we observe that extreme granularities (e.g., $P = 4$ or $P = 32$) lead to higher performance variability across runs, whereas $P = 16$ maintains consistent numerical stability. To ensure visibility of these behavioral shifts despite subtle absolute differences, the y-axis scales for MSE and MAE are independently adjusted (e.g., in Figure 8a, MSE spans $[1.255, 1.275]$); however, the exact numerical optima are reported alongside the plots to prevent misinterpretation.

These findings further validate the design of the PIE module, demonstrating that regional aggregation at an appropriate scale is critical for balancing noise suppression and detail preservation. By establishing a moderate regional receptive field, PiTransformer successfully circumvents the instability of point-wise methods while avoiding the catastrophic representation loss associated with over-coarsening in volatile temporal manifolds.

4.10. Training Stability and Statistical Convergence Analysis

To quantitatively evaluate the reliability of the forecasting results and address the concern regarding weight initialization luck, we define the training stability through the sample standard deviation (σ) across multiple independent trials:

$$\sigma = \sqrt{\frac{1}{N-1}\sum_{i=1}^{N}(E_i - \bar{E})^2} \tag{11}$$

where $N = 5$ is the number of trials, E_i represents the prediction error (MSE) of the i-th trial, and $\bar{E}$ is the mean error. By employing the Bessel's correction ($N - 1$), we provide an unbiased estimation of the numerical stability, which is essential for assessing the reproducibility of models in volatile financial manifolds. The statistical results for all evaluated models are summarized in Table 8.

Table 8. Statistical stability analysis across five independent trials ($itr = 5$). Metrics denote Mean $\pm$ Std of MSE. **Bold and underlined** values indicate the best performance, and underlined values represent the second-best results.

Dataset	DLinear	iTransformer	PatchTST	SegRNN	PiTransformer (Ours)
SPY	1.4425 ± 0.0035	1.3849 ± 0.0111	1.2823 ± 0.0066	2.2592 ± 0.0053	$\mathbf{1.2594 \pm 0.0013}$
Exchange	0.1028 ± 0.0007	0.1075 ± 0.0014	0.0899 ± 0.0008	$\mathbf{0.0864 \pm 0.0012}$	0.0940 ± 0.0019

As evidenced in Table 8, on the high-entropy SPY dataset, PiTransformer achieves the most consistent performance with an order-of-magnitude reduction in standard deviation ($\sigma = 0.0013$) compared to its structural baseline, iTransformer ($\sigma = 0.0111$). This significant tightening of the error distribution suggests that the gated patch-wise architecture facilitates a more stable optimization trajectory, effectively anchoring the model against stochastic initialization variance. While DLinear and PatchTST also exhibit reasonable stability, they fail to reach the lower error floor established by PiTransformer.

Notably, the performance of SegRNN reveals a distinct "mean-variance paradox" in financial data: while it achieves the best results on the smoother Exchange dataset, its MSE on the volatile SPY benchmark is substantially higher (2.2592). This suggests a structural vulnerability to extreme non-stationarity, where a low standard deviation (0.0053) indicates that the model is consistently converging to a sub-optimal representation. In contrast, PiTransformer maintains superior accuracy and high precision (± 0.0013) simultaneously. These quantitative findings confirm that the observed performance gains are not artifacts of initialization luck but are rooted in the model's inherent capacity to prune stochastic noise and extract stable temporal features.

5. Conclusions

In this study, we have formalized and evaluated the PiTransformer, a re-parameterized forecasting framework engineered to resolve the inherent tension between localized temporal fidelity and global multivariate coupling. By synthesizing patch-wise regional representations with a gated inverted attention mechanism, the proposed architecture successfully circumvents the representation limitations that frequently compromise the predictive stability of vanilla Transformer variants in high-entropy regimes.

Through systematic empirical evaluations across heterogeneous domains—ranging from stochastic financial markets to periodic physical systems—the following key insights are summarized:

1. **Efficacy of Patch-wise Inversion**: The integration of patch-based segments into the inverted manifold provides a measurable defense against representation collapse. Our results, particularly the 8.9% MSE reduction on the SPY benchmark, confirm that regional tokenization effectively preserves fine-grained temporal structures that are typically attenuated by the global point-wise projections utilized in standard inversion schemes [10].
2. **Adaptive Filtering and Domain Bifurcation via VTG**: The Variable–Temporal Gating (VTG) module functions as a learnable informational distillation layer inspired by the Information Bottleneck principle. The observed performance bifurcation—where significant gains are concentrated in low-SNR environments while marginal trade-offs occur in deterministic signals—indicates that the gating mechanism acts as an adaptive filter. This observation suggests that while VTG is a distinctive tool for chaotic manifolds, its filtering intensity must be carefully modulated to avoid suppressing useful periodic patterns in high-SNR physical regimes, a direction we leave for future research.
3. **Empirical Interpretability**: Analysis of the learned attention manifold demonstrates that PiTransformer reveals interpretable interaction patterns, such as the lead-lag relationship between *Volume* and *Close* tokens. This transition from black-box regression toward a more transparent interaction structure provides an essential layer of evidence-based modeling for practical forecasting applications.
4. **Architectural and Computational Economy**: By maintaining a parsimonious architecture of approximately 53.3K parameters, PiTransformer achieves a competitive balance between model capacity and efficiency. While maintaining a highly efficient $O((V \cdot N)^2 \cdot d_{model})$ computational complexity (where V is the number of variables and N is the number of patches), our model reaches the performance echelon of more complex state-of-the-art architectures, including PatchTST [9] and SegRNN [13].

Future research may explore various noteworthy extensions to further refine the theoretical and empirical boundaries of the proposed framework.

First, the development of *adaptive multi-scale tokenization* strategies represents a pivotal direction for capturing the multi-resolution temporal dynamics inherent in financial manifolds. Given the fractal nature of market volatility, replacing fixed-length patches with learnable segments could improve the model's sensitivity to heterogeneous temporal regimes, allowing for a more nuanced extraction of both high-frequency micro-structures and low-frequency macro-trends.

Second, integrating gated inversion logic with *pre-trained time-series foundation models* [16] offers a viable path for enhancing cross-domain zero-shot transferability. In this context, the VTG module could be re-parameterized as a "noise-aware adapter" to bridge the gap between general-purpose temporal representations and domain-specific stochasticity, leveraging large-scale historical knowledge while maintaining specialized filtering capabilities.

Third, the fusion of *multimodal stochastic information* offers an avenue for enriching the inverted representation space. Future iterations of PiTransformer could incorporate exogenous unstructured data, such as market sentiment indices, through a unified cross-variable attention manifold. This would facilitate the modeling of complex "event-to-price" propagation mechanisms, providing a more comprehensive understanding of non-stationary market shifts within a unified interaction space.

Finally, evaluating the framework's efficacy in *real-time execution environments* will provide deeper insights into its practical utility. Transitioning from passive statistical

forecasting to active decision-making—where PiTransformer acts as a robust state-encoder within a reinforcement learning loop—could effectively bridge the gap between predictive fidelity and optimal resource allocation in increasingly chaotic environments.

Author Contributions: Conceptualization, L.Z. and K.C.; methodology, K.C.; software, K.C.; validation, K.C.; formal analysis, K.C.; investigation, K.C.; resources, L.Z.; data curation, K.C.; writing—original draft preparation, L.Z. and K.C.; writing—review and editing, L.Z.; visualization, K.C.; supervision, L.Z.; project administration, L.Z.; funding acquisition, L.Z. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by Research on Online Deep Transfer Learning of Time Series based on Financial Data (Grant No.: Qian Ke He Ji Chu [2024] general 520) supported by Guizhou Province Basic Research Plan (Natural Sciences) Projects. Additionally, it received funding from the Sixth Batch of Gui-Zhou Province High-level Innovative Talent Training Program (Grant No.: Zhu Ke He Tong [2022]011) and the Guiyang University New Degree Awarding Point Cultivation and Construction Project in 2025 [Gyxk202502].

Data Availability Statement: The data presented in this study are openly available. The ETT datasets (ETTh1 and ETTm1) and the Exchange Rate dataset can be found in the Time-Series-Library repository at https://github.com/thuml/Time-Series-Library (accessed on 28 December 2025). The SPY dataset was obtained from Yahoo Finance (https://finance.yahoo.com/, accessed on 28 December 2025).

Acknowledgments: The authors would like to express their sincere gratitude to the School of Electronic Information Engineering of Guiyang University for providing a favorable research environment for promoting the integration of FinTech and deep learning methods. In addition, they would like to express their sincerest gratitude to every researcher who provided assistance in this research.

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A.N.; Kaiser, Ł.; Polosukhin, I. Attention Is All You Need. In Proceedings of the 31st Annual Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA, 4–9 December 2017; Volume 30, pp. 5998–6008. Available online: https://proceedings.neurips.cc/paper_files/paper/2017/hash/3f5ee243547dee91fbd053c1c4a845aa-Abstract.html (accessed on 20 March 2026).
2. 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, Virtual, 19–21 May 2021; Volume 35, pp. 11106–11115. [CrossRef]
3. Wu, H.; Xu, J.; Wang, J.; Long, M. Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. In Proceedings of the 35th Conference on Neural Information Processing Systems (NeurIPS 2021), Virtual, 6–14 December 2021; Volume 34, pp. 22419–22430. Available online: https://proceedings.neurips.cc/paper_files/paper/2021/hash/bcc0d400288793e8bdcd7c19a8ac0c2b-Abstract.html (accessed on 20 March 2026).
4. Zhou, T.; Ma, Z.; Wen, Q.; Wang, X.; Sun, L.; Jin, R. Fedformer: Frequency enhanced decomposed transformer for long-term series forecasting. In Proceedings of the 39th International Conference on Machine Learning (ICML 2022), Baltimore, MD, USA, 17–23 July 2022; Volume 162, pp. 27268–27286. Available online: https://proceedings.mlr.press/v162/zhou22g.html (accessed on 20 March 2026).
5. Zeng, A.; Chen, M.; Zhang, L.; Xu, Q. Are Transformers Effective for Time Series Forecasting? In Proceedings of the AAAI Conference on Artificial Intelligence, Washington, DC, USA, 7–14 February 2023; Volume 37, pp. 12845–12853. [CrossRef]
6. Liu, Y.; Wu, H.; Wang, J.; Long, M. Non-stationary Transformers: Exploring the Stationarity in Time Series Forecasting. In Proceedings of the 36th Conference on Neural Information Processing Systems (NeurIPS 2022), New Orleans, LA, USA, 28 November–9 December 2022; Volume 35, pp. 19393–19406. Available online: https://proceedings.neurips.cc/paper_files/paper/2022/hash/4054556fcaa934b0bf76da52cf4f92cb-Abstract-Conference.html (accessed on 20 March 2026).
7. Wu, H.; Hu, T.; Liu, Y.; Zhou, H.; Wang, J.; Long, M. TimesNet: Temporal 2D-Variation Modeling for General Time Series Analysis. In Proceedings of the 11th International Conference on Learning Representations (ICLR 2023), Kigali, Rwanda, 1–5 May 2023. Available online: https://openreview.net/forum?id=ju_Uqw384Oq (accessed on 20 March 2026).

8. Kitaev, N.; Kaiser, Ł.; Levskaya, A. Reformer: The efficient transformer. In Proceedings of the 8th International Conference on Learning Representations (ICLR 2020), Addis Ababa, Ethiopia (Virtual), 26 April–1 May 2020. Available online: https://openreview.net/forum?id=rkgNKkHtvB (accessed on 20 March 2026).
9. Nie, Y.; Nguyen, N.H.; Sinthong, P.; Kalagnanam, J. A Time Series is Worth 64 Words: Long-term Forecasting with Transformers. In Proceedings of the 11th International Conference on Learning Representations (ICLR 2023), Kigali, Rwanda, 1–5 May 2023. Available online: https://openreview.net/forum?id=oigW0ASMNz (accessed on 20 March 2026).
10. Liu, Y.; Hu, T.; Zhang, H.; Wu, H.; Wang, S.; Ma, L.; Long, M. iTransformer: Inverted Transformers Are Effective for Time Series Forecasting. In Proceedings of the 12th International Conference on Learning Representations (ICLR 2024), Vienna, Austria, 7–11 May 2024. Available online: https://openreview.net/forum?id=JePfAI8fah (accessed on 20 March 2026).
11. Lim, B.; Zohren, S. Time-series forecasting with deep learning: A survey. *Philos. Trans. A Math. Phys. Eng. Sci.* **2021**, *379*, 20200209. [CrossRef] [PubMed]
12. Wang, S.; Wu, H.; Shi, X.; Hu, T.; Luo, H.; Ma, L.; Zhang, J.Y.; Zhou, J. Timemixer: Decomposable multiscale mixing for time series forecasting. In Proceedings of the 12th International Conference on Learning Representations (ICLR 2024), Vienna, Austria, 7–11 May 2024. Available online: https://openreview.net/forum?id=yFvr4eRT0W (accessed on 20 March 2026).
13. Lin, S.; Lin, W.; Wu, W.; Zhao, F.; Mo, R.; Zhang, H. SegRNN: Segment recurrent neural network for long-term time series forecasting. In Proceedings of the 12th International Conference on Learning Representations (ICLR 2024), Vienna, Austria, 7–11 May 2024. Available online: https://openreview.net/forum?id=66T7p9OAnS (accessed on 20 March 2026).
14. Wang, Y.; Wu, H.; Dong, J.; Qin, G.; Zhang, H.; Liu, Y.; Qiu, Y.; Wang, J.; Long, M. Timexer: Empowering transformers for time series forecasting with exogenous variables. In Proceedings of the 38th Conference on Neural Information Processing Systems (NeurIPS 2024), Vancouver, BC, Canada, 10–15 December 2024; Volume 37, pp. 469–498. [CrossRef]
15. Zhang, T.; Zhang, Y.; Cao, W.; Bian, J.; Yi, X.; Zheng, S.; Li, J. Less is more: Fast multivariate time series forecasting with light sampling-oriented MLP structures. *arXiv* **2022**, arXiv:2207.01186. [CrossRef]
16. Zhou, T.; Niu, P.; Wang, X.; Sun, L.; Jin, R. One Fits All: Power General Time Series Analysis by Pretrained LM. In Proceedings of the Conference on Neural Information Processing Systems (NeurIPS 2023), New Orleans, LA, USA, 10–16 December 2023; Volume 36, pp. 43357–43373. Available online: https://proceedings.neurips.cc/paper_files/paper/2023/hash/86c17de05579cde52025f9984e6e2ebb-Abstract-Conference.html (accessed on 20 March 2026).
17. Oreshkin, B.N.; Carpov, D.; Chapados, N.; Bengio, Y. N-BEATS: Neural basis expansion analysis for interpretable time series forecasting. In Proceedings of the 8th International Conference on Learning Representations (ICLR 2020), Addis Ababa, Ethiopia, 26–30 April 2020; Volume 107, pp. 1–12. Available online: https://openreview.net/forum?id=r1ecqn4YwB (accessed on 20 March 2026).
18. Challu, C.; Olivares, K.G.; Oreshkin, B.N.; Ramirez, F.G.; Canseco, M.M.; Dubrawski, A. Nhits: Neural hierarchical interpolation for time series forecasting. In Proceedings of the 37th AAAI Conference on Artificial Intelligence, Washington, DC, USA, 7–14 February 2023; Volume 37, pp. 6989–6997. [CrossRef]
19. Xu, Z.; Zeng, A.; Xu, Q. FITS: Modeling time series with 10k parameters. In Proceedings of the 12th International Conference on Learning Representations (ICLR 2024), Vienna, Austria, 7–11 May 2024. Available online: https://openreview.net/forum?id=m2getD1hpk (accessed on 20 March 2026).
20. Zhang, Y.; Yan, J. Crossformer: Transformer utilizing cross-dimension dependency for multivariate time series forecasting. In Proceedings of the 11th International Conference on Learning Representations (ICLR 2023), Kigali, Rwanda, 1–5 May 2023. Available online: https://openreview.net/forum?id=vSVLM2j9eie (accessed on 20 March 2026).
21. Wen, Q.; Zhou, T.; Zhang, C.; Chen, W.; Ma, Z.; Yan, J.; Sun, L. Transformers in time series: A survey. In Proceedings of the 32nd International Joint Conference on Artificial Intelligence (IJCAI 2023), Macao, China, 19–25 August 2023; pp. 6750–6760. [CrossRef]
22. Lim, B.; Arık, S.Ö.; Loeff, N.; Pfister, T. Temporal fusion transformers for interpretable multi-horizon time series forecasting. *Int. J. Forecast.* **2021**, *37*, 1748–1764. [CrossRef]
23. Kim, T.; Kim, J.; Tae, Y.; Park, C.; Choi, J.H.; Choo, J. Reversible instance normalization for accurate time-series forecasting against distribution shift. In Proceedings of the 10th International Conference on Learning Representations (ICLR 2022), Virtual, 25–29 April 2022. Available online: https://openreview.net/forum?id=cGDakQo1C0p (accessed on 20 March 2026).
24. Kingma, D.P.; Ba, J. Adam: A Method for Stochastic Optimization. In Proceedings of the 3rd International Conference on Learning Representations (ICLR 2015), San Diego, CA, USA, 7–9 May 2015. Available online: https://openreview.net/forum?id=8gmWwjFyLj (accessed on 20 March 2026).