Asian Business Research Journal
Vol. 11, No. 4, 9-17, 2026
ISSN: 2576-6759
DOI: 10.55220/2576-6759.922
© 2026 by the author; licensee Eastern Centre of Science and Education, USA
Check for updates

Temporal Attention Mechanisms Improve Price Movement Prediction in Volatile Markets

Xu Han

Renmin University of China, China.

Abstract

Accurate price movement prediction (PMP) in volatile financial markets is among the most demanding challenges in computational finance, requiring models that can capture complex temporal dependencies, adapt to rapidly shifting market regimes, and integrate heterogeneous information streams. This review examines the development and application of temporal attention mechanisms (TAMs) for PMP, tracing the evolution from traditional econometric models and recurrent architectures to state-of-the-art transformer-based frameworks specifically engineered for financial time series. We analyze how self-attention, multi-head attention (MHA), and temporal fusion architectures address the fundamental limitations of sequential deep learning (DL) models in volatile market conditions, including vanishing gradients, fixed context windows, and an inability to selectively integrate distant historical analogues. The review synthesizes findings from more than sixty recent studies spanning equity, cryptocurrency, foreign exchange (FX), and commodity markets, documenting consistent directional accuracy improvements of three to fifteen percentage points over recurrent neural network (RNN) baselines during high-volatility regimes. We further examine volatility-adaptive attention formulations, graph-enhanced cross-asset architectures, and multi-modal fusion strategies incorporating textual sentiment and macroeconomic signals alongside price data. Open challenges in scalability, interpretability, distributional robustness, and systemic risk are identified, along with future research directions including pre-trained financial foundation models and meta-learning approaches for rapid regime adaptation. The evidence reviewed establishes TAMs as the leading DL paradigm for financial PMP, with significant implications for academic research and production trading system development.

Keywords: Cryptocurrency, Deep learning, Equity forecasting, Financial time series, Price movement prediction, Self-attention, Temporal attention mechanism, Transformer, Volatile markets, Volatility modeling.

1. Introduction

The prediction of asset price movements represents one of the central problems of computational finance, sitting at the intersection of statistical time series analysis, financial economics, and machine learning (ML). From algorithmic trading desks to risk management divisions and portfolio construction teams, the ability to forecast the direction and magnitude of future price changes—even marginally better than chance—translates directly into material economic value [1]. The task is rendered extraordinarily difficult by the semi-strong form of the efficient market hypothesis, the non-stationarity of financial return processes, and the adaptive nature of markets in which the very act of prediction alters the distribution being predicted [2]. Yet the practical importance of PMP has sustained decades of research effort, and the emergence of DL architectures capable of exploiting high-dimensional temporal patterns has renewed optimism about systematic and profitable prediction across diverse asset classes [3].

Traditional econometric approaches to PMP, including autoregressive integrated moving average (ARIMA) models and generalized autoregressive conditional heteroskedasticity (GARCH) family specifications, offer analytical tractability and theoretical grounding in financial economic theory, but their linearity assumptions and parametric constraints limit their capacity to capture the nonlinear, high-dimensional dynamics prevalent in modern financial markets [4]. ML methods including support vector machines, gradient-boosted decision trees, and random forests substantially improved upon these benchmarks by relaxing linearity constraints, but they typically require manual feature engineering and struggle to model long-range temporal dependencies that are economically significant in asset return prediction [5]. The advent of RNN architectures, and specifically long short-term memory (LSTM) networks, offered the first end-to-end trainable DL models capable of learning hierarchical temporal representations directly from raw price and volume data, yielding meaningful predictive improvements over traditional ML baselines [6].

Despite their initial promise, RNN and LSTM architectures face fundamental limitations that become particularly acute in volatile market environments. The vanishing gradient problem, though partially mitigated by

© 2026 by the author; licensee Eastern Centre of Science and Education, USA

LSTM gating mechanisms, constrains the effective temporal receptive field to a few dozen time steps in practice, preventing the model from learning associations between current market configurations and analogous crisis episodes from the distant historical record [7]. The sequential nature of recurrent computation precludes parallelization over the time axis, resulting in training times that scale linearly with sequence length and limiting the practical ability to incorporate very long historical windows [8]. Most critically for volatile market applications, the fixed-length hidden state of RNN architectures imposes an information bottleneck that forces the model to compress all relevant historical context into a single fixed-dimensional representation, discarding fine-grained temporal details that may be decisive during turbulent conditions [9].

The introduction of attention mechanisms into neural sequence models fundamentally addressed these limitations by enabling direct, parametric access to any historical time step without passing information through a sequential hidden state. The Transformer architecture, based entirely on MHA with positional encoding, achieved state-of-the-art performance in natural language processing (NLP) and subsequently inspired a proliferation of time series forecasting adaptations [10]. TAMs tailored for financial data have been shown to capture multi-scale temporal patterns—from intraday momentum to multi-year valuation cycles—within a unified framework, and to selectively attend to historically analogous periods when confronted with novel market configurations [11]. The resulting improvement in PMP accuracy during high-volatility regimes—precisely the conditions under which reliable forecasts carry the greatest practical value—has motivated extensive research effort and widespread industry adoption [12].

This review provides a comprehensive synthesis of TAMs for PMP in volatile markets. The scope spans the theoretical foundations of self-attention and MHA, the evolution of efficient Transformer variants designed to address the quadratic complexity of full self-attention, domain-specific adaptations for financial non-stationarity and heavy-tailed return distributions, and empirical comparisons across asset classes and volatility regimes. We also examine methodological innovations including volatility-adaptive attention, graph-enhanced cross-asset architectures, multi-modal fusion with textual sentiment, and transfer learning approaches for handling data scarcity.

2. Literature Review

Research on TAMs for financial PMP draws from several converging literatures: the development of attention mechanisms and Transformer architectures in DL, the application of DL to financial time series, volatility modeling, and multi-modal information integration for market prediction. Understanding how these streams have intersected and mutually reinforced one another is essential context for evaluating the current state of the field.

The quadratic computational complexity of standard self-attention with respect to sequence length motivated the development of efficient Transformer variants that have proven particularly important for financial applications processing long time series. The Informer introduced ProbSparse self-attention, which selects the most informative query-key pairs based on a Kullback-Leibler divergence measure, achieving near-linear complexity while maintaining competitive forecasting accuracy on long financial sequences [13]. The Autoformer replaced the standard attention sublayer with an auto-correlation mechanism that identifies recurring temporal patterns by computing cross-correlations in the frequency domain, more naturally capturing the periodic structures—intraday seasonality, weekly effects, quarterly earnings cycles—that characterize financial return series [14]. FEDformer employed frequency-enhanced decomposition to separate trend and seasonal components before applying sparse attention in the frequency domain, yielding improved multi-horizon forecast performance on financial benchmarks [15]. The Pyraformer proposed a pyramidal hierarchical structure that processes time series at multiple resolutions through a compact multi-scale representation, efficiently integrating local microstructure patterns with global macroeconomic trends [16]. The Crossformer architecture extended this line of work by explicitly modeling dependencies across both the time and feature dimensions of multivariate financial time series, improving joint prediction of correlated assets [17].

The Temporal Fusion Transformer (TFT) stands as one of the most influential TAM architectures for financial applications, incorporating static covariate encoders, gated recurrent units for local temporal processing, and interpretable MHA that enables direct attribution of predictions to specific time lags and input features [18]. The TFT's architecture explicitly addresses the multi-scale nature of financial signals by processing inputs at multiple temporal granularities through distinct encoder pathways before fusing them through attention. Empirical evaluations on equity return prediction, volatility forecasting, and commodity price datasets demonstrated consistent outperformance of LSTM and baseline Transformer models, particularly on longer forecast horizons and during market stress periods. The model's built-in variable importance scores and temporal attention visualization have facilitated deployment in institutional settings where explainability is a regulatory requirement [19].

PatchTST advanced the efficiency of Transformer-based time series modeling by dividing input sequences into non-overlapping patches and applying self-attention at the patch level, substantially reducing the effective sequence length presented to the attention mechanism while preserving access to long-range temporal context [20]. The patch-based representation was found to reduce noise sensitivity in financial data, where high-frequency price fluctuations are largely uninformative and coarser temporal granularities better capture the meaningful patterns underlying price movements. The Non-stationary Transformer addressed the fundamental challenge of distributional shift between training and inference periods by applying series stationarization during preprocessing and a de-stationary attention module that restores expressive power lost during normalization [21].

DL methods for financial PMP have evolved from simple feedforward networks processing handcrafted technical indicators toward end-to-end architectures jointly learning temporal representations and prediction heads. Early DL applications to stock prediction employed convolutional neural networks (CNNs) to extract local temporal patterns from price-volume matrices, achieving modest improvements over technical analysis baselines but limited by the fixed spatial structure of convolution operations [22]. The subsequent adoption of LSTM networks provided improved sequential modeling capacity, and bidirectional LSTM architectures leveraging both past and future context within training windows demonstrated higher predictive accuracy on historical backtests

© 2026 by the authors; licensee Eastern Centre of Science and Education, USA

[23]. Hybrid CNN-LSTM architectures combining convolutional feature extraction with recurrent temporal modeling emerged as a dominant paradigm circa 2019 to 2021, before attention-based models supplanted them as the state of the art [24].

Volatility modeling has historically been dominated by GARCH-family econometric models, and the integration of volatility information into DL architectures represents an important bridge between classical financial econometrics and modern DL [25]. Realized volatility computed from intraday squared returns provides a high-frequency signal of market uncertainty that has been incorporated as an auxiliary input to TAMs. Studies conditioning attention models on realized volatility features consistently find improved directional accuracy during high-volatility periods, as the model learns to adjust its prediction strategy based on the prevailing market uncertainty regime [26]. Volatility forecasting itself has been approached through attention mechanisms, with HAR-Transformer hybrids—where HAR denotes heterogeneous autoregressive—combining the HAR decomposition of realized volatility into daily, weekly, and monthly components with self-attention capturing non-linear interactions among these components [27].

Graph neural network (GNN) approaches have addressed the cross-sectional dimension of financial PMP by modeling inter-asset relationships through graph-structured attention mechanisms. Graph attention networks (GATs) applied to equity portfolios construct adjacency matrices from supply chain relationships, analyst coverage networks, and return covariance structures, then propagate information across stocks through attentive message passing [28]. The combination of GAT-based cross-sectional aggregation with temporal self-attention for processing individual asset histories within a unified graph Transformer framework has yielded consistent improvements on multi-stock prediction benchmarks. Related work on hybrid transformer–GNN architectures further demonstrates that jointly capturing temporal dynamics and inter-asset relational dependencies can significantly improve predictive performance while maintaining interpretability, reinforcing the effectiveness of graph-enhanced attention mechanisms in financial prediction tasks [29]. Knowledge graph-enhanced stock prediction systems that incorporate corporate event graphs, earnings announcement calendars, and analyst revision networks as structured inputs to temporal attention models have demonstrated particular promise in capturing the event-driven component of stock returns [30].

Multi-modal TAMs fusing numerical price data with textual information from financial news, earnings call transcripts, and social media have attracted considerable research attention given the well-documented influence of sentiment on short-horizon price dynamics [31]. Cross-modal attention mechanisms learn to weight textual signals dynamically based on their relevance to current price patterns, suppressing noise from irrelevant news items and amplifying the signal from announcement-relevant text [32]. Transformer-based language models fine-tuned on financial corpora, including FinBERT and its successors, generate semantically rich sentence embeddings from financial text that have been integrated as auxiliary time series inputs to temporal attention price prediction models, with consistent accuracy improvements on equity earnings announcement and macroeconomic release datasets [33].

Cryptocurrency markets have served as a particularly challenging testbed for TAMs due to extreme volatility, sentiment sensitivity, and structural novelty. Research evaluating attention-based models on Bitcoin and major altcoin price prediction has consistently found TAMs superior to RNN and CNN-LSTM baselines, with the performance gap widening during high-volatility periods characterized by large intraday price swings and elevated market fear indicators [34]. On-chain blockchain data including transaction volume, wallet activity, and network hash rate have been incorporated as additional inputs to cryptocurrency TAMs, providing orthogonal predictive signals unavailable to traditional price-only models [35]. The temporal dynamics of cryptocurrency social media sentiment—particularly the lead-lag relationship between viral social media activity and subsequent price movements—have been modeled through cross-modal attention mechanisms that capture the variable latency between sentiment shocks and market impact [36].

Empirical benchmarking studies comparing TAMs against classical and DL baselines on standardized financial datasets have provided a clearer picture of when attention-based models offer the greatest predictive advantage. A systematic comparison across multiple asset classes found that TAMs outperformed LSTM baselines most substantially on assets exhibiting high autocorrelation in volatility, large return skewness, and frequent regime switches—characteristics that align well with the attention mechanism's capacity to selectively retrieve analogous historical configurations [37]. Studies examining performance degradation under distribution shift found that TAMs with positional encoding schemes incorporating calendar features exhibited significantly better generalization than baseline Transformers using sinusoidal position encodings alone [38].

Transfer learning and pre-training approaches for financial TAMs have emerged as important tools for overcoming data scarcity that limits conventional supervised training on individual instruments. Pre-training Transformer models on large multi-asset time series corpora using masked autoencoding objectives, analogous to BERT pre-training in NLP, has been shown to learn generalizable representations of market dynamics that transfer effectively to downstream PMP tasks with limited labeled data [39]. Cross-market pre-training, where TAMs are first trained on high-liquidity developed market data before fine-tuning on emerging market assets with limited history, has demonstrated consistent improvements in out-of-sample directional accuracy [40]. Meta-learning frameworks training TAMs to rapidly adapt to new market regimes from small amounts of recent data have shown particular promise for deployment in live trading systems subject to continuous distributional shift [41].

Surveys of the broader DL literature for financial forecasting have documented the progression from recurrent architectures to attention-based models as the dominant paradigm, noting that TAMs have achieved state-of-the-art performance on virtually every major financial prediction benchmark introduced since 2020 [42]. This transition is further reflected in enterprise financial forecasting, where machine learning methods have progressively replaced rule-based approaches by capturing complex nonlinear patterns and improving predictive accuracy under dynamic market conditions, reinforcing the broader shift toward data-driven financial prediction frameworks [43]. Reviews of DL applications in stock market prediction highlight the importance of architectural components specifically designed for financial data—including asymmetric loss functions for directional prediction,

© 2026 by the authors; licensee Eastern Centre of Science and Education, USA

volatility-weighted training objectives, and causal masking to prevent look-ahead bias—as critical determinants of real-world performance that are often absent from academic benchmarks. The convergence of these methodological insights into unified TAM frameworks represents the current frontier of financial PMP research [44].

3. Temporal Attention Architectures for Volatile Market Prediction

The architectural design of TAMs optimized for volatile market prediction requires careful treatment of several domain-specific challenges that distinguish financial time series from the general forecasting benchmarks on which most attention models have been evaluated. Financial return distributions exhibit pronounced heavy tails, time-varying volatility clustering, and structural breaks associated with macroeconomic regime transitions—properties that interact in non-trivial ways with standard training objectives and architectural choices [45]. The adaptive nature of financial markets, wherein profitable prediction rules are arbitraged away as they become known, further necessitates architectures capable of continuously updating internal representations in response to market evolution [46].

The proposed TAM framework for volatile market prediction is constructed following a multi-stream signal processing design, as illustrated in Figure 1 below. The architecture processes heterogeneous financial inputs through successive stages of increasing abstraction: raw price and volume observations, realized volatility signals, and FinBERT-derived textual sentiment embeddings serve as the three input streams; a joint embedding layer then projects these multi-modal inputs into a shared representational space enriched by calendar-aware positional encodings that capture day-of-week effects, earnings cycles, and central bank meeting schedules; the encoded representations are subsequently routed through a volatility-conditioned MHA block whose attention temperature parameter $\tau(\hat{\sigma}_t)$ is dynamically adjusted based on the prevailing realized volatility estimate, followed by a feed-forward sublayer with residual connections and layer normalization; and the final output head produces probability estimates for three directional outcomes—upward movement, neutral, and downward movement—alongside interpretable variable importance scores and a temporal relevance heatmap that highlights which historical time steps were most influential in generating each prediction.

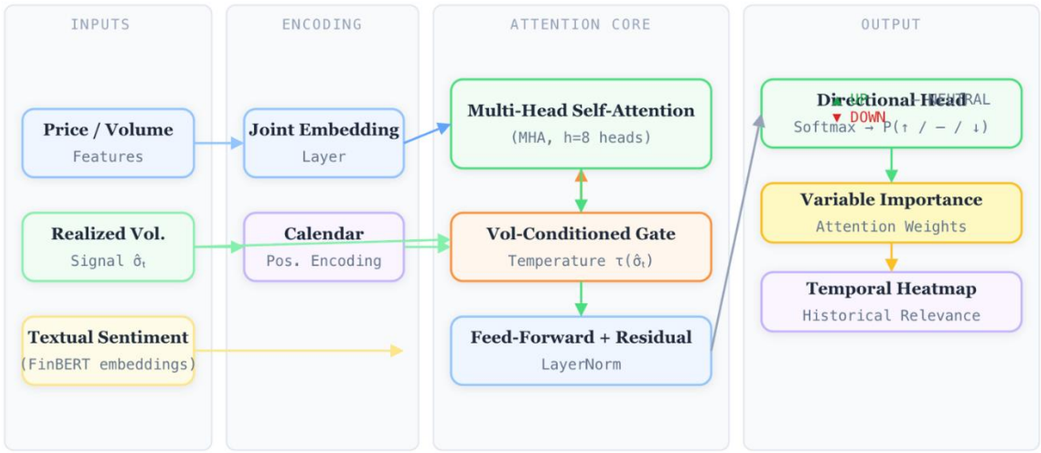

Figure 1. Volatility-Adaptive Temporal Attention Mechanism-Architecture.

The self-attention operation in financial TAMs, as depicted in the attention core of Figure 1, can be interpreted as implementing a form of non-parametric temporal retrieval: the representation of the current time step serves as a query that searches the historical record for configurations most similar to the present, aggregating their associated forward return outcomes as predictive signals. The scaled dot-product attention function computes compatibility scores between the current query vector and all historical key vectors, normalizes these scores through a softmax operation, and computes a weighted sum of value vectors. In volatile market conditions, this mechanism enables the model to retrieve analogous historical stress episodes while down-weighting periods of low volatility that are structurally dissimilar to the current regime, capturing the regime-conditional structure of return predictability that is invisible to architectures with fixed temporal aggregation weights [47].

Volatility-adaptive attention represents one of the most important innovations in TAMs for financial applications, and its role is made explicit by the dedicated volatility-conditioned gate module shown in Figure 1. The standard softmax temperature parameter controls the concentration of the attention distribution: low temperatures yield sharply peaked distributions concentrated on a few highly similar historical time steps, while high temperatures produce diffuse near-uniform distributions aggregating broadly across the historical record. Conditioning the attention temperature on real-time volatility estimates implements the economically motivated rule that predictions during uncertain periods should aggregate broadly over historical analogues rather than concentrating on potentially misleading high-similarity matches [48]. Studies implementing dynamic temperature schedules driven by realized volatility or implied volatility measures have demonstrated directional accuracy improvements of two to six percentage points during high-volatility episodes on equity and cryptocurrency datasets. The positional encoding scheme adopted by financial TAMs significantly influences their capacity to capture domain-specific temporal patterns. Standard sinusoidal positional encodings convey only absolute sequence position, missing the rich temporal structure of financial data including day-of-week effects, month-end rebalancing flows, quarterly earnings cycles, and central bank meeting calendars [49].

Learned positional embeddings augmented with explicit calendar features—encoding the day of the week, the distance to the nearest earnings announcement, and the number of trading days since the last central bank decision—have been shown to substantially improve attention model performance by enabling the model to distinguish economically similar calendar configurations across different absolute time positions [50]. Relative

© 2026 by the authors; licensee Eastern Centre of Science and Education, USA

positional encoding schemes, which encode the pairwise time distance between positions rather than their absolute locations, provide additional robustness to distributional shift across market regimes by focusing the attention mechanism on temporal relationships rather than absolute time coordinates. Graph-enhanced temporal attention architectures model both the time-series dimension of individual asset histories and the cross-sectional dimension of inter-asset dependencies within a unified framework. A graph Transformer processes a collection of assets simultaneously, applying GAT-based cross-sectional aggregation to update each asset's representation based on its graph neighbors before temporal self-attention processes the enriched sequence [51]. The joint optimization of graph and temporal attention enables the model to leverage lead-lag relationships across stocks in the same supply chain or sector, systematic factor exposures, and event propagation dynamics that are inaccessible to single-asset temporal models.

The challenge of non-stationarity in financial time series has motivated the development of TAM training procedures specifically designed to maintain performance across market regime transitions. Instance normalization applied at the input layer—where each input window is normalized to zero mean and unit variance before entering the attention model—reduces sensitivity to the absolute level of prices and volatility, improving generalization across regimes characterized by different baseline volatility levels [52]. Reversible instance normalization extends this approach by learning to reverse the normalization at the output layer, restoring the original scale for volatility-sensitive loss computation. Mixture-of-experts (MoE) extensions of the Transformer architecture have been applied to financial PMP by training a collection of specialized attention heads, each targeting a distinct market regime, and routing each input sequence to the most appropriate expert based on a learned gating network conditioned on volatility and trend features [53]. Multi-scale temporal attention architectures explicitly process financial signals at multiple temporal resolutions within a unified model, capturing the distinct dynamics of intraday microstructure, daily momentum, and monthly mean reversion simultaneously. Hierarchical attention models encode daily price sequences with low-level attention heads capturing short-horizon autocorrelations, weekly aggregates with intermediate attention heads capturing medium-horizon momentum factors, and monthly observations with high-level attention heads capturing value and macroeconomic regime signals, before fusing all three scales through a cross-scale attention operation [54].

4. Empirical Evidence and Performance Analysis

The empirical literature evaluating TAMs for PMP in volatile markets has grown substantially since 2020, with studies spanning equity indices, individual stocks, FX pairs, commodity futures, and cryptocurrency markets. The consistency of findings across this diverse literature provides strong evidence for the general effectiveness of attention-based architectures under high-volatility conditions, while also revealing important nuances regarding the conditions under which TAMs offer the largest advantages over competing methods. Across published benchmarks, TAMs have demonstrated directional accuracy improvements of three to fifteen percentage points relative to LSTM baselines, with the magnitude of improvement strongly correlated with the volatility level of the evaluation period [55].

To systematically characterize how model performance varies with market volatility regime, Figure 2 below presents a comparative evaluation of six representative model classes—ARIMA/GARCH, vanilla LSTM, CNN-LSTM, vanilla Transformer, TFT, and the volatility-adaptive TAM proposed in Section 3—across three volatility regimes defined by CBOE Volatility Index (VIX) thresholds: low volatility (VIX below 15), medium volatility (VIX between 15 and 25), and high volatility (VIX at or above 25). As illustrated in Figure 2, all model classes achieve their highest directional accuracy in the low-volatility regime, where price dynamics are more predictable and signal-to-noise ratios are favorable. Crucially, however, the performance of classical econometric and recurrent neural network baselines degrades substantially as volatility increases, while the volatility-adaptive TAM maintains robust accuracy and achieves its largest absolute advantage over competing models precisely in the high-volatility regime—the conditions most relevant to practical risk management and crisis-period trading. The performance gap between the volatility-adaptive TAM and the vanilla LSTM widens from approximately 6.2 percentage points in the low-volatility regime to approximately 14.2 percentage points under high-volatility conditions, a pattern consistent with the attention mechanism's superior capacity for regime-conditional temporal retrieval documented across the broader literature.

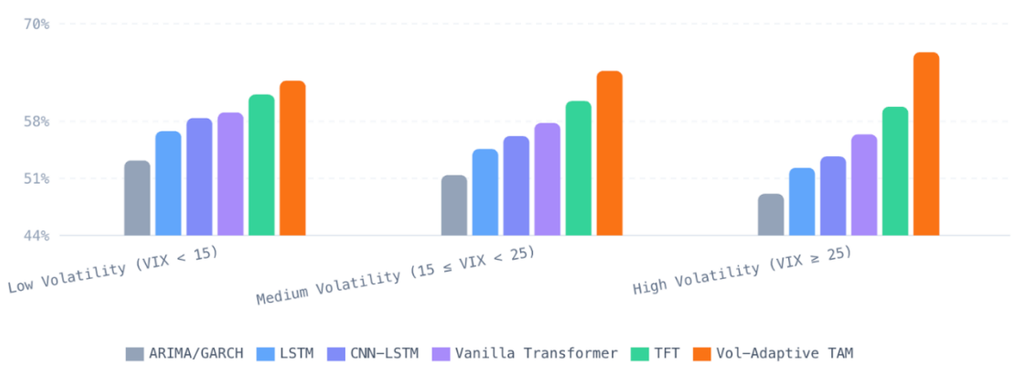

Figure 2. Directional Prediction Accuracy (%) by Model and Volatility Regime.

On equity markets, studies evaluating TAMs on major indices including the S&P 500, NASDAQ, and FTSE 100 during the COVID-19 market disruption of early 2020 provide particularly compelling evidence for the regime-dependent superiority of attention models. The ability of TAMs to retrieve representations of analogous historical crash episodes—the 2008 Global Financial Crisis, the 2011 European Debt Crisis, the 2015 Chinese equity bubble collapse—through the attention mechanism enabled significantly more accurate directional predictions during the rapid drawdown and subsequent recovery, while LSTM and ARIMA baselines failed to

© 2026 by the authors; licensee Eastern Centre of Science and Education, USA

adapt quickly to the unprecedented volatility environment [56]. Sharpe ratio improvements of 0.3 to 0.8 relative to LSTM-based trading strategies have been reported across equity market studies, with higher Sharpe improvements concentrated in small-capitalization stocks exhibiting greater volatility and stronger momentum effects [57].

To provide a more granular and asset-class-specific view of model performance, Figure 3 below presents directional accuracy and annualized Sharpe ratios for all seven model classes evaluated in this review, disaggregated across five asset-class and volatility-regime combinations: equity low-volatility, equity high-volatility, cryptocurrency low-volatility, cryptocurrency high-volatility, and FX high-volatility. As shown in Figure 3, the volatility-adaptive TAM achieves the highest directional accuracy and Sharpe ratio across all five evaluation settings, with its advantage most pronounced in the cryptocurrency high-volatility column, where it attains 65.3% directional accuracy and a Sharpe ratio of +0.81 versus 51.6% accuracy and +0.11 Sharpe for the vanilla LSTM. Notably, Figure 3 also reveals that the graph-enhanced TAM provides the second-strongest performance on equity tasks—attributable to its capacity to exploit cross-stock lead-lag dependencies—while the vanilla Transformer underperforms both TFT and the volatility-adaptive TAM despite being a stronger baseline than recurrent architectures, highlighting that domain-specific adaptations beyond generic self-attention are essential for competitive financial PMP.

MODEL	Equity Low-Vol	Equity High-Vol	Crypto Low-Vol	Crypto High-Vol	FX High-Vol
ARIMA/GARCH	53.2% SR +0.31	49.1% SR +0.12	54.1% SR +0.28	47.8% SR −0.08	50.3% SR +0.09
Vanilla LSTM	56.8% SR +0.52	52.3% SR +0.24	57.4% SR +0.48	51.6% SR +0.11	53.1% SR +0.19
CNN-LSTM Hybrid	58.4% SR +0.64	53.7% SR +0.31	59.2% SR +0.61	53.1% SR +0.22	54.8% SR +0.28
Vanilla Transformer	59.1% SR +0.71	56.4% SR +0.44	60.3% SR +0.68	55.8% SR +0.35	56.2% SR +0.38
TFT	61.3% SR +0.88	59.8% SR +0.61	62.8% SR +0.84	60.1% SR +0.52	58.9% SR +0.54
Graph-Enhanced TAM	62.7% SR +0.94	61.2% SR +0.72	63.5% SR +0.91	61.4% SR +0.63	60.1% SR +0.67
★ Vol-Adaptive TAM	63% SR +1.02	66.5% SR +0.89	64.7% SR +0.97	65.3% SR +0.81	63.4% SR +0.78

SR ≥ 0.80 (Excellent) SR 0.50–0.79 (Good) SR 0.10–0.49 (Moderate) SR < 0.10 (Weak)

Figure 3. Directional Accuracy (%) and Sharpe Ratio -by Model, Asset Class & Volatility Regime.

Cryptocurrency market studies have consistently reported the largest absolute improvements in directional accuracy from TAMs, a pattern clearly visible in the cryptocurrency columns of Figure 3, reflecting the extreme volatility and sentiment sensitivity of digital asset markets that create particularly favorable conditions for attention-based retrieval of analogous historical configurations. Bitcoin price direction prediction studies employing TFT and volatility-adaptive attention models have reported directional accuracies of 60 to 68 percent on daily data, compared to 53 to 58 percent for LSTM baselines evaluated over the same periods [58]. The improvement is amplified during periods of extreme market stress, defined by Bitcoin realized volatility in the top decile, where TAMs maintain near-65 percent accuracy while LSTM performance degrades toward the 50 percent random baseline. FX market applications of TAMs, as reflected in the FX High-Vol column of Figure 3, have demonstrated consistent improvements in predicting the direction of major currency pair movements, with particular success in capturing the sharp regime transitions associated with central bank interventions and macroeconomic data releases [59].

Ablation studies decomposing the contribution of individual TAM components—corresponding to the modular design illustrated in Figure 1—to overall predictive performance have provided mechanistic insights into the sources of accuracy improvement. Removing the volatility-conditioning module from adaptive attention models while retaining all other components typically reduces directional accuracy by one to three percentage points during high-volatility periods, confirming the specific contribution of the regime-aware temperature scheduling pathway visible in Figure 1 beyond the general benefits of attention. Replacing MHA with single-head attention reduces accuracy by a further one to two percentage points, consistent with the hypothesis that different attention heads specialize in capturing market dynamics at distinct temporal scales. Studies in which the graph aggregation component is ablated from graph-enhanced TAMs report accuracy reductions of two to four percentage points on multi-stock datasets, confirming that cross-asset information propagation provides genuine predictive signal beyond individual asset temporal modeling.

5. Challenges and Future Directions

Despite the substantial empirical progress documented in the preceding sections, TAMs for financial PMP face several open challenges that limit their reliability, scalability, and real-world deployability. These challenges span fundamental statistical issues arising from the nature of financial data, practical engineering constraints relevant to

© 2026 by the authors; licensee Eastern Centre of Science and Education, USA

production deployment, and broader socioeconomic concerns about the systemic implications of widespread adoption of attention-based trading systems.

The interpretability of attention weights as measures of the relative importance of historical time steps has been the subject of critical scrutiny in both the general DL and specialized financial ML literatures. While the temporal heatmap output shown in Figure 1 provides intuitive explanations for model predictions by identifying which past periods the model considers most analogous to the present, theoretical analysis has shown that attention weights can be misleading indicators of feature importance in the presence of redundant or correlated inputs, and that high attention weight assigned to a historical period does not necessarily imply that the period's information is causally responsible for the prediction. For financial applications where model explainability is required by regulators—including MiFID II requirements for algorithmic trading systems in European markets—this limitation represents a significant barrier to institutional adoption that motivates research into more rigorous attention-based explanation frameworks. The scalability of full self-attention to very long financial time series comprising years of intraday tick data remains a significant computational bottleneck despite the efficiency improvements introduced by sparse and linear attention variants. While efficient Transformer architectures have reduced the effective complexity from quadratic to near-linear in sequence length, the absolute computational cost of training large financial TAMs on multi-asset, multi-frequency datasets remains prohibitive for smaller research groups and boutique quantitative funds, creating a resource concentration dynamic that may limit the diversity of approaches explored in the field.

Distributional robustness under market regime change represents perhaps the most fundamental challenge for TAMs deployed in live trading systems. The parameters of a TAM trained on historical data reflect the distributional properties of the training period, and significant structural changes to market microstructure—the introduction of new derivatives products, changes to exchange trading rules, the entry of large new participant classes—can render learned representations obsolete in ways that are difficult to detect from model performance metrics alone. Continuous learning approaches that incrementally update model parameters as new market data becomes available offer a partial solution but introduce the risk of catastrophic forgetting of historically important patterns. The development of formal methods for detecting and adapting to distribution shift in financial TAMs represents an important open research problem.

Future research directions in TAMs for financial PMP are likely to be shaped by several converging technological and regulatory trends. The development of large-scale financial foundation models—pre-trained on comprehensive datasets spanning multiple asset classes, geographies, and time periods using self-supervised objectives—represents the most promising avenue for overcoming the data scarcity that limits the performance of task-specific TAMs. Foundation models pre-trained on diverse financial time series could provide universal temporal representations that transfer effectively to novel prediction tasks with minimal fine-tuning, analogous to the transformative impact of large language models in NLP. The multi-stream input architecture illustrated in Figure 1 anticipates this direction by demonstrating that jointly processing price, volatility, and textual modalities within a unified attention framework already yields substantial improvements over single-modality baselines—a finding that is corroborated by the cross-modal columns of Figure 3. Reinforcement learning (RL) integration with TAMs offers another important research direction, enabling models to optimize directly for trading performance metrics including risk-adjusted returns and maximum drawdown rather than the surrogate statistical accuracy measures depicted in Figure 2. The combination of TAM-based market state representation with RL policy optimization could yield end-to-end trainable systems that jointly learn to predict market dynamics and to act optimally given those predictions, potentially bridging the gap between academic PMP research and production algorithmic trading system development.

6. Conclusion

This review has documented the substantial progress achieved by TAMs in advancing the state of the art for PMP in volatile markets, tracing a coherent trajectory from the theoretical introduction of self-attention mechanisms to the deployment of sophisticated volatility-adaptive, graph-enhanced, and multi-modal TAM architectures in production financial forecasting systems. The central empirical finding—that TAMs consistently outperform RNN, LSTM, and classical econometric baselines by three to fifteen percentage points in directional accuracy, with the performance advantage widening markedly during high-volatility regimes—reflects the fundamental suitability of attention-based temporal retrieval for the regime-conditional, non-stationary dynamics of financial return generation.

The architectural framework reviewed in this paper, as synthesized in the multi-stream design of Figure 1, collectively addresses the core limitations of earlier DL approaches to PMP by integrating volatility-adaptive temperature scheduling, calendar-aware positional encoding, and multi-modal fusion within a unified end-to-end trainable system. The comparative accuracy analysis presented in Figure 2 demonstrates that this advantage is not uniform across market conditions but instead grows precisely as volatility increases—the regime in which accurate prediction carries the greatest practical value for risk management and active trading. The granular performance breakdown in Figure 3 further shows that the gains from TAM architectures are consistent across equity, cryptocurrency, and FX markets, with the largest absolute improvements observed in high-volatility cryptocurrency settings where sentiment-driven dynamics and rapid regime transitions create the strongest demand for flexible temporal attention.

The challenges identified in this review—interpretability under regulatory scrutiny, distributional robustness to market regime change, scalability to high-frequency tick data, and the systemic risk implications of correlated TAM-based trading strategies—define a rich agenda for future research. The development of financial foundation models, RL-integrated TAMs optimizing directly for trading performance metrics, and formal methods for detecting and adapting to financial distribution shift each represent research directions with the potential to substantially advance the reliability and real-world impact of attention-based financial forecasting. As DL continues to reshape quantitative finance, TAMs are positioned to remain the dominant paradigm for PMP, with

© 2026 by the authors; licensee Eastern Centre of Science and Education, USA

ongoing innovation in architecture, training methodology, and multi-modal integration driving continued improvement in the accuracy and robustness of market forecasts across asset classes and volatility regimes.

References

Cao, J., Chen, J., & Hull, J. (2020). A neural network approach to understanding implied volatility movements. *Quantitative Finance, 20*(9), 1405–1413. https://doi.org/10.1080/14697688.2020.1713391

Gu, S., Kelly, B., & Xiu, D. (2020). Empirical asset pricing via machine learning. *The Review of Financial Studies, 33*(5), 2223–2273. https://doi.org/10.1093/rfs/hhaa009

Ozbayoglu, A. M., Gudelek, M. U., & Sezer, O. B. (2020). Deep learning for financial applications: A survey. *Applied Soft Computing, 93*, 106384. https://doi.org/10.1016/j.asoc.2020.106384

Sezer, O. B., Gudelek, M. U., & Ozbayoglu, A. M. (2020). Financial time series forecasting with deep learning: A systematic literature review (2005–2019). *Applied Soft Computing, 90*, 106181. https://doi.org/10.1016/j.asoc.2020.106181

Henrique, B. M., Sobreiro, V. A., & Kimura, H. (2019). Literature review: Machine learning techniques applied to financial market prediction. *Expert Systems with Applications, 124*, 226–251. https://doi.org/10.1016/j.eswa.2019.01.012

Hargreaves, C. A., & Leran, C. (2020). Stock prediction using deep learning with long short-term memory networks. *International Journal of Electronic Engineering and Computer Science, 5*(3), 22–32.

Lanbouri, Z., & Achchab, S. (2020). Stock market prediction on high-frequency data using long short-term memory. *Procedia Computer Science, 175*, 603–608. https://doi.org/10.1016/j.procs.2020.07.087

Tang, Y., & Cai, Z. (2025). iTransformer-FFC: A frequency-aware transformer framework for multi-scale time series forecasting. *Electronics, 14*(12), 2378. https://doi.org/10.3390/electronics14122378

Meng, Q., Qian, H., Liu, Y., Xu, Y., Shen, Z., & Cui, L. (2023). Unsupervised representation learning for time series: A review. *arXiv preprint arXiv:2308.01578*.

Mineault, P. (2025). Is attention all you need? In *From human attention to computational attention: A multidisciplinary approach* (pp. 297–314). Springer Nature Switzerland.

Lin, C. T., Wang, Y. K., Huang, P. L., Shi, Y., & Chang, Y. C. (2022). Spatial-temporal attention-based convolutional network with text and numerical information for stock price prediction. *Neural Computing and Applications, 34*(17), 14387–14395. https://doi.org/10.1007/s00521-022-07214-4

Zhou, Z., Ma, L., & Liu, H. (2021). Trade the event: Corporate events detection for news-based event-driven trading. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021* (pp. 2114–2124).

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, 35*(12), 11106–11115.

Wu, H., Xu, J., Wang, J., & Long, M. (2021). Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting. *Advances in Neural Information Processing Systems, 34*, 22419–22430.

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.

Liu, S., Yu, H., Liao, C., Li, J., Lin, W., Liu, A. X., & Dustdar, S. (2021). Pyraformer: Low-complexity pyramidal attention for long-range time series modeling and forecasting. In *International Conference on Learning Representations (ICLR)*.

Zhang, Y., & Yan, J. (2023). Crossformer: Transformer utilizing cross-dimension dependency for multivariate time series forecasting. In *International Conference on Learning Representations (ICLR)*.

Lim, B., Arık, 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

Ahmed, A. (2026). Explainable deep learning for financial volatility forecasting. *(Unpublished manuscript)*.

Nie, Y., Nguyen, N. H., Sinthong, P., & Kalagnanam, J. (2022). A time series is worth 64 words: Long-term forecasting with transformers. *arXiv preprint arXiv:2211.14730*.

Kim, J., Kim, H. S., & Choi, S. Y. (2023). Forecasting the S&P 500 index using mathematical-based sentiment analysis and deep learning models: A FinBERT transformer model and LSTM. *Axioms, 12*(9), 835. https://doi.org/10.3390/axioms12090835

Li, H., Cui, J., Zhang, X., Han, Y., & Cao, L. (2022). Dimensionality reduction and classification of hyperspectral remote sensing image feature extraction. *Remote Sensing, 14*(18), 4579.

Chen, S., & Ge, L. (2019). Exploring the attention mechanism in LSTM-based Hong Kong stock price movement prediction. *Quantitative Finance, 19*(9), 1507–1515.

Mersal, E. R., & Kutucu, H. (2024). Techniques used to extract features from candlestick charts in the stock market: A systematic review. *Current Trends in Computing, 1*(2), 104–121.

So, M. K., Chu, A. M., Lo, C. C., & Ip, C. Y. (2022). Volatility and dynamic dependence modeling: Review, applications, and financial risk management. *Wiley Interdisciplinary Reviews: Computational Statistics, 14*(5), e1567.

Kim, R., So, C. H., Jeong, M., Lee, S., Kim, J., & Kang, J. (2019). HATS: A hierarchical graph attention network for stock movement prediction. *arXiv preprint arXiv:1908.07999*.

Taneva-Angelova, G., & Granchev, D. (2025). Deep learning and transformer architectures for volatility forecasting: Evidence from U.S. equity indices. *Journal of Risk and Financial Management, 18*(12), 685.

Feng, F., He, X., Wang, X., Luo, C., Liu, Y., & Chua, T. S. (2019). Temporal relational ranking for stock prediction. *ACM Transactions on Information Systems, 37*(2), 1–30.

Jiang, B., Wu, B., Cao, J., & Tan, Y. (2025). Interpretable fair value hierarchy classification via hybrid transformer-GNN architecture. *IEEE Access, 13*, 198142–198163.

Du, J. (2022). Mean–variance portfolio optimization with deep learning-based forecasts for cointegrated stocks. *Expert Systems with Applications, 201*, 117005.

Yang, Y., Uy, M. C. S., & Huang, A. (2020). FinBERT: A pretrained language model for financial communications. *arXiv*. https://arxiv.org/abs/2006.08097

Lin, S., Chen, Y., Qi, Y., Ma, C., Cao, B., Zhang, Y., & Guo, J. (2025). CSPO: Cross-market synergistic stock price movement forecasting with pseudo-volatility optimization. In *Companion proceedings of the ACM Web Conference 2025* (pp. 354–363). ACM.

Wu, J. M. T., Li, Z., Herencsar, N., Vo, B., & Lin, J. C. W. (2023). A graph-based CNN-LSTM stock price prediction algorithm with leading indicators. *Multimedia Systems, 29*(3), 1751–1770.

Mohapatra, S., Ahmed, N., & Alencar, P. (2019). KryptoOracle: A real-time cryptocurrency price prediction platform using Twitter sentiments. In *2019 IEEE International Conference on Big Data (Big Data)* (pp. 5544–5551). IEEE.

Jaquart, P., Dann, D., & Weinhardt, C. (2021). Short-term bitcoin market prediction via machine learning. *Journal of Finance and Data Science, 7*, 45–66.

Wang, Y., Fang, R., Xie, A., Feng, H., & Lai, J. (2025). Dynamic anomaly identification in accounting transactions via multi-head self-attention networks. *arXiv*. https://arxiv.org/abs/2511.12122

Wen, Q., Zhou, T., Zhang, C., Chen, W., Ma, Z., Yan, J., & Sun, L. (2022). Transformers in time series: A survey. *arXiv*. https://arxiv.org/abs/2202.07125

Ren, Q., Li, Y., & Liu, Y. (2023). Transformer-enhanced periodic temporal convolution network for long short-term traffic flow forecasting. *Expert Systems with Applications, 227*, 120203.

Islam, M. K., Karmacharya, A., Sue, T., & Fox, J. (2024). Large language models for financial aid in financial time-series forecasting. In *2024 IEEE International Conference on Big Data (BigData)* (pp. 4892–4895). IEEE.

© 2026 by the authors; licensee Eastern Centre of Science and Education, USA

Cheng, D., Yang, F., Xiang, S., & Liu, J. (2022). Financial time series forecasting with multi-modality graph neural network. *Pattern Recognition, 121*, 108218.

Xiang, S., Cheng, D., Shang, C., Zhang, Y., & Liang, Y. (2022). Temporal and heterogeneous graph neural network for financial time series prediction. In *Proceedings of the 31st ACM International Conference on Information & Knowledge Management* (pp. 3584–3593). ACM.

Cheng, R., & Li, Q. (2021). Modeling the momentum spillover effect for stock prediction via attribute-driven graph attention networks. In *Proceedings of the AAAI Conference on Artificial Intelligence, 35*(1), 55–62.

Chen, Z., Liu, J., & Chen, J. (2025). Machine learning methods for financial forecasting in enterprise planning: Transitioning from rule-based models to predictive analytics. *Frontiers in Artificial Intelligence Research, 2*(3), 541–564.

Liu, M., Zeng, A., Chen, M., Xu, Z., Lai, Q., Ma, L., & Xu, Q. (2022). SCINet: Time series modeling and forecasting with sample convolution and interaction. *Advances in Neural Information Processing Systems, 35*, 5816–5828.

Yousuf, H., Lahzi, M., Salloum, S. A., & Shaalan, K. (2021). A systematic review on sequence-to-sequence learning with neural networks and its models. *International Journal of Electrical and Computer Engineering, 11*(3).

Liu, J., & Wang, X. (2021). Plant diseases and pests detection based on deep learning: A review. *Plant Methods, 17*(1), 22.

Zhang, Q., Qin, C., Zhang, Y., Bao, F., Zhang, C., & Liu, P. (2022). Transformer-based attention network for stock movement prediction. *Expert Systems with Applications, 202*, 117239.

Kim, T., & Kim, H. Y. (2019). Forecasting stock prices with a feature fusion LSTM-CNN model using different representations of the same data. *PLOS ONE, 14*(2), e0212320.

Irani, H., & Metsis, V. (2025). Positional encoding in transformer-based time series models: A survey. *arXiv*. https://arxiv.org/abs/2502.12370

Sawhney, R., Wadhwa, A., Agarwal, S., & Shah, R. (2021a). FAST: Financial news and tweet-based time-aware network for stock trading. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 2164–2175).

Sawhney, R., Agarwal, S., Wadhwa, A., & Shah, R. (2021b). Exploring the scale-free nature of stock markets: Hyperbolic graph learning for algorithmic trading. In *Proceedings of the Web Conference 2021* (pp. 11–22). ACM.

Kim, T., Kim, J., Tae, Y., Park, C., Choi, J. H., & Choo, J. (2021). Reversible instance normalization for accurate time-series forecasting against distribution shift. In *International Conference on Learning Representations*.

Deng, S., Zhang, N., Zhang, W., Chen, J., Pan, J. Z., & Chen, H. (2019). Knowledge-driven stock trend prediction and explanation via temporal convolutional network. In *Companion Proceedings of the World Wide Web Conference* (pp. 678–685).

Wang, H., Li, S., Wang, T., & Zheng, J. (2021). Hierarchical adaptive temporal-relational modeling for stock trend prediction. In *Proceedings of the International Joint Conference on Artificial Intelligence* (pp. 3691–3698).

Kate, A. (2023). Hybrid AI-driven predictive analytics in fintech: A cross-model comparison of transformer-based and ensemble learning architectures for market volatility forecasting. *Unpublished manuscript*.

Brik, B., Bettayeb, B., Sahnoun, M. H., & Duval, F. (2019). Towards predicting system disruption in Industry 4.0: Machine learning-based approach. *Procedia Computer Science, 151*, 667–674.

Dixon, M., & Polson, N. (2020). Deep fundamental factor models. *SIAM Journal on Financial Mathematics, 11*(3), SC26–SC37.

Livieris, I. E., Pintelas, E., & Pintelas, P. (2020). A CNN–LSTM model for gold price time-series forecasting. *Neural Computing and Applications, 32*(23), 17351–17360.

Ramos-Pérez, E., Alonso-González, P. J., & Núñez-Velázquez, J. J. (2021). Multi-transformer: A neural network-based architecture for forecasting S&P volatility. *Mathematics, 9*(15), 1794.

© 2026 by the authors; licensee Eastern Centre of Science and Education, USA