Temporal Fusion Transformers for S&P 500 Return Forecasting with Mixed-Frequency Macroeconomic Data

Sam Ehrle
Arizona State University
sehrle@asu.edu

Gourishankar Mahadeo Bansode
Arizona State University
gourishankar@asu.edu

Ojas Makarand Deodhar
Arizona State University
odeodhar@asu.edu

Aman Pandey
Arizona State University
apand105@asu.edu

Danyal Khorami
Arizona State University
dkhorami@asu.edu

Abstract

Predicting stock returns is challenging due to low signal-to-noise ratios, regime shifts between market conditions, and mixed-frequency data where features update at different rates. We evaluate Temporal Fusion Transformers (TFT) for S&P 500 return forecasting using daily market indicators alongside monthly macroeconomic releases, comparing against ARIMAX and LSTM baselines. Our experiments reveal systematic prediction collapse where models converge to constant outputs despite low validation loss. Analysis shows encoder representations correctly detect regime shifts while output layers fail to translate these patterns into predictions. Staleness-aware modifications in both input and attention space worsen the collapse, while regime-aware attention with learned volatility gates shows interpretable behavior but limited generalization. Quantile loss proves anti-correlated with directional accuracy, requiring model selection by prediction diversity. We find that temporal aggregation enables signal extraction: daily multi-horizon and weekly single-step both achieve approximately 2% excess directional accuracy, suggesting both access the same underlying signal through different aggregation mechanisms.

1. Introduction

Predicting the market is a considerable challenge, primarily due to the fundamental properties of how prices fluctuate. At its core, the efficient market hypothesis, formulated by Fama et al. (1970) [1], states that prices in capital markets fully reflect all available information, rendering financial markets "efficient" in the sense that we can not systematically exploit historical data and publicly known signals for excess returns. On the one hand, the random walk model is confirmed to be natural in light of the independence and identical distribution of asset prices, with returns being unpredictable based solely on prices.

In terms of financial market forecasting, any signal that can go beyond the 50% level of accuracy demonstrated by a random walk process can be considered significant in light of the inherent uncertainty of efficient markets. For instance, signals with even minimal advantages over the random walk, such as 52% accuracy, can be considered outstanding and may represent an opportunity for exploiting market inefficiencies. The efficient market theory proposed by Fama states that signals significantly deviating from the random walk model are unlikely to persist in the market.

Fama and others have documented that financial return data is just plain noisy. Economic news, the mindset of investors, and technological changes constantly and easily move prices and may obscure the signal. This is why most models tend to underperform simple statistical benchmarks. Identifying a potential signal is not enough. The actual problem is to extract a stable, implementable signal from all the random noise, particularly when the data is uncertain. Markets also show regime changes where model parameters shift between bullish and bearish periods [2, 3].

The main problem in time-series forecasting lies in capturing non-local temporal structures, the distant dependencies that conventional sequential models (like RNNs) often miss. The transformer-based architecture is becoming one of the state-of-the-art solutions, as it utilizes a self-attention mechanism, allowing for changes in weights based on historical observations and providing a framework to model complex, non-sequential relationships. This breakthrough has made it a leading approach, exemplified by influential models like the Temporal Fusion Transformer (TFT) [4]. TFT specifically addresses multi-horizon forecasting with heterogeneous inputs, utilizing attention to intelligently blend historical lags, external covariates, and con-

temporaneous signals—a necessity in mixed-frequency environments typical of financial prediction. Besides TFT, the Autoformer's decomposition framework for better periodic modeling [5] and Informer's sparse self-attention for efficiency in long sequences [6], provide a strong dominance in transformer-based architecture models for time series forecasting. These reasons make the Transformer-based architecture one of the choices for modeling S&P 500 returns forecasting with mixed-frequency macroeconomic data.

Non-stationarity is a fundamental challenge in financial prediction. Regime changes, shifts in trends, and structural breaks in market data mean that historical patterns most often fail to forecast the future behavior of the market. This characteristic creates unstable dynamics where statistical properties of the data change over time. Recent work addresses non-stationarity through adapting segmentation of the input series [7].

Another challenge in predicting the stock market is data heterogeneity. This problem arises when features update at different frequencies, as some update daily (VIX volatility) while others update monthly or even quarterly (CPI inflation, GDP growth). The mixed-frequency in such data is not predictable by standard processing methods; for instance, forward-fill propagates the most recent value across time periods and, regardless of data freshness, treats 30-day-old CPI measurements identically to yesterday's stock price. Such an approach overlooks the temporal decay of information content, eventually creating an asymmetry where stale macroeconomic data receives equal weighting to fresh market signals. Ghysels, Santa-Clara, and Valkanov (2004) note that "the relevant information is high frequency data, whereas the variable of interest is sampled at a lower frequency," and introduced Mixed-Data Sampling (MIDAS) regressions to address this challenge through polynomial-weighting schemes that parsimoniously aggregate high-frequency observations using distributed lag polynomials[8].

This work makes the following contributions: (1) Systematic characterization of TFT failure modes on mixed-frequency financial data, including prediction collapse where models converge to constant positive outputs, and encoder-output gradient disconnect. (2) Evaluation of staleness-aware mechanisms in input space (staleness features) and attention space (learned penalties), showing universal performance degradation. (3) Development and evaluation of regime-conditioned attention with learned volatility gates. (4) Identification of quantile loss anti-correlation with directional accuracy, establishing prediction diversity as the appropriate checkpoint selection criterion. (5) Demonstration that temporal aggregation through multi-horizon prediction or weekly frequency achieves approximately 2% excess directional accuracy.

2. Related Work

2.1. Deep Learning for Financial Forecasting

To deal with these limitations, many of the researchers have increasingly turned to neural network approaches, most notably, LSTM networks. Fischer et al. (2018) [9] were among the first to use LSTM architectures in predicting the direction of movements and daily returns for constituents of S&P 500 from 1992-2015. They demonstrated that LSTMs have better generalization accuracy and performance per unit than methods that do not make use of memory. The immediate advantage of the LSTM was due to its processing of a sequential input, 240-day return sequences, and the capture of complex temporal dependencies through the memory cell architecture comprised of forget, input, and output gates, which control the flow of information across time steps. Notably, their models captured complex short-term reversal patterns, volatility clustering, high-beta stock characteristics, and regime shifts beyond the reach of traditional statistical methods.

Much of the deep learning research focuses only on price histories. It overlooks macroeconomic factors. Our work extends these approaches using attention-based architectures, aiming to capture better temporal dependencies and macro-market relationships than both classical and recurrent baselines.

2.2. Transformer Architectures for Time Series

Transformer architectures have been introduced as a strong alternative to recurrent models for time series forecasting. However, the usage comes with significant computational and architectural challenges that need to be addressed. For instance, the quadratic complexity of a standard transformer restricts the application of long sequences, and its point-wise attention operation may not necessarily be the most effective way to interact with the temporal nature of time series data. In order to address the issues, Lim et al. (2021) proposed the Temporal Fusion Transformer (TFT) that utilizes the power of recurrent layers to model local temporal patterns and applies self-attention to capture long-range dependencies. The variable selection network of the model dynamically selects different inputs; hence, it can manage mixed input types without requiring manual feature specification [4].

On the other hand, Zhou et al. (2021) brought up the Informer, which is equipped with a ProbSparse attention mechanism that locates the most informative queries for a selective focus to reduce the complexity to $O(L \log L)$ and also allows generative predictions for the entire forecast horizon; therefore, inference is significantly faster for long-term forecasting [6]. Wu et al. (2021) addressed different patterns and seasonal components of time series by directly integrating decomposition into the Autoformer ar-

chitecture and using an Auto-Correlation mechanism to accurately capture periodic patterns [5]. We adapt the TFT method to the financial constraints.

2.3. Mixed-Frequency Data

Mixed-frequency data challenges occur when high-frequency variables hold important information for a target that is observed at a lower frequency [8]. MIDAS regression solves this issue by using polynomial weighting schemes to efficiently combine high-frequency observations into low-frequency predictions without averaging everything [8]. We build on MIDAS's polynomial weighting concept by applying it to attention. This lets the model learn the temporal weights instead of depending on preset functional forms.

2.4. Data Quality Issues in Financial Machine Learning

The IFC (2025) asserts, "poor data quality can lead to unreliable models, which may impede decision-making and prediction," when handling "data sparsity, noisy data, and dynamic environments" [10]. More traditional preprocessing techniques, like forward-fill, solve ignored data problems by perpetually carrying forward the last available observation for every empty time space. The method gives equal weight to a macro release from two months ago as to a real-time observation of market volatility for predicting market direction. Ignoring the time reduction of informational value results in an imbalance where outdated economic indicators are weighted as heavily as current financial signals, which systematically divides prediction errors [10].

2.5. Attention Mechanisms in Finance

Applications of temporal fusion transformers (TFTs) in financial markets show their use for stock price forecasting and predicting market trends [11, 12]. Recent studies also look at attention-based models. These models combine structured price data and unstructured news sentiment to better classify financial market movements, like the price direction of the FTSE 100 [13].

3. Methods

3.1. Data Pipeline

Our dataset spans January 1991 to October 2025, comprising daily S&P 500 returns and mixed-frequency predictor variables. Features include daily market indicators (the CBOE Volatility Index (VIX), 10-year Treasury yield, 10Y-2Y yield spread) and monthly macroeconomic releases (CPI-derived inflation with a roughly 14-day publication lag). VIX measures the expected 30-day S&P 500 volatility and is commonly used as a proxy to characterize market regimes.

Macroeconomic data presents a look-ahead bias risk. Indicators may be revised post-release, and naive implementations are at risk of using values unavailable at prediction time. We address this through vintage date alignment using the ALFRED database [14], retrieving only data vintages available on each historical date.

Data is partitioned chronologically: training (1991-2015), validation (2015-2020), and test (2020-2025). This ensures evaluation spans distinct market regimes, with validation including the COVID-19 disruption and test covering the post-pandemic period and 2022-2023 Fed tightening cycle (Figure 1).

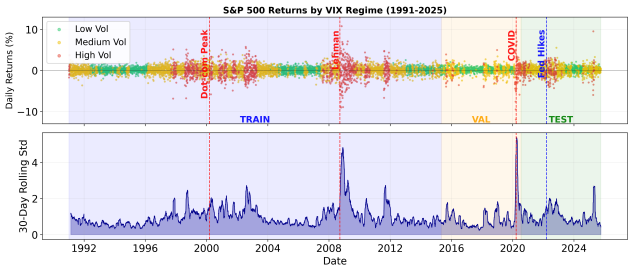

Figure 1. S&P 500 daily returns colored by VIX regime (green: low volatility, yellow: medium, red: high) with 30-day rolling volatility below. Background shading indicates train/validation/test splits. Major market events (dot-com peak, Lehman collapse, COVID-19, Fed rate hikes) cluster with high-volatility regimes, motivating regime-aware modeling approaches.

3.2. Model Architecture

We employ the Temporal Fusion Transformer (TFT) [4], which combines LSTM encoders for local temporal processing with interpretable multi-head attention for long-range dependencies. The model is trained with quantile loss across seven quantiles (0.02, 0.1, 0.25, 0.5, 0.75, 0.9, 0.98) with additional experiments using 3 and 5 quantiles. We test prediction horizons $h \in \{1, 3, 5, 10, 20, 30\}$ using TFT's native multi-horizon decoder. Multi-horizon performance is reported as the average directional accuracy across all horizons from 1 to h. We use 2 attention heads, a hidden dimension of 16, a hidden continuous size of 16, and a dropout of 0.1. The encoder observes 20 historical time steps. Training uses the Ranger optimizer (RAdam + Lookahead) with a learning rate of 5e-4, ReduceLROnPlateau scheduler (patience=4), and a gradient clipping of 0.1. Checkpoints are selected by validation prediction standard deviation rather than validation loss.

For baseline comparison, we implement ARIMAX [15] and a standard LSTM on the same one-day-ahead S&P 500 return prediction task. Both models are trained on the fixed-alignment `core_proposal` daily dataset using the same set of exogenous covariates and a strictly chronological 70/15/15 train–validation–test split. The LSTM takes

as input sliding windows of the past 15 trading days and consists of a three-layer LSTM with 128 hidden units, 0.2 dropout, and two fully connected layers mapping the final hidden state to the next-day return. We optimize mean-squared error using Adam (learning rate 10^{-3}, batch size 64) and select the checkpoint with the lowest validation loss. For ARIMAX, we conduct a grid search over $(p, d, q) \in \{0, \ldots, 3\} \times \{0, \ldots, 2\} \times \{0, \ldots, 3\}$ based on the univariate return series' AIC, then refit the selected ARIMAX(p, d, q) model with the same exogenous regressors on the training set and report its performance on the held-out test period.

3.3. Architectural Modifications

We extend the base TFT architecture with several modifications targeting the failure modes identified in our experiments. Figure 2 illustrates the overall system, showing the dual-head output structure and regime-aware attention mechanism.

Loss Function Penalties. We extend the standard quantile loss with regularization terms to prevent common predictive failure modes. A directional diversity penalty discourages unidirectional predictions by means of penalizing batches where more than 90% of predictions share the same sign:

$$\mathcal{L}_{div} = \lambda_{Div} \cdot \max(0, \hat{b} - \tau)^2 \tag{1}$$

where $\hat{b}$ is the observed directional bias and $tau = 0.9$ is the threshold. This regularization is based on empirical analysis showing real market returns never exceed this threshold over 30-day windows. Additional penalties target minimum prediction variance (anti-collapse) and temporal consistency between sequential predictions. These modifications are implemented by subclassing pytorch-forecasting's QuantileLoss [16].

Regime-Conditional Output Layer. Inspired by mixture-of-experts (MoE) architectures [17], we replace the TFT output layer with an MoE architecture, where parallel expert heads specialize in different market regimes. A router network assigns samples to experts based on either learned hidden state features or deterministic VIX thresholds. This is intended to test whether regime-specific prediction heads can improve performance when the model detects regime shifts in attention patterns but fails to adapt the output behavior accordingly.

Classification Head. Following multi-task approaches [18], we add a parallel classification head to the TFT encoder output, trained jointly with quantile regression via a combined loss:

$$\mathcal{L}_{total} = \alpha \mathcal{L}_{quantile} + \beta \mathcal{L}_{CE} \tag{2}$$

This serves as a diagnostic tool to assess if the classification head learns while regression fails, indicating that the encoder representation is informative, but the regression output layer cannot exploit it. We evaluate both direction prediction (binary up/down) and regime classification (VIX-based volatility states).

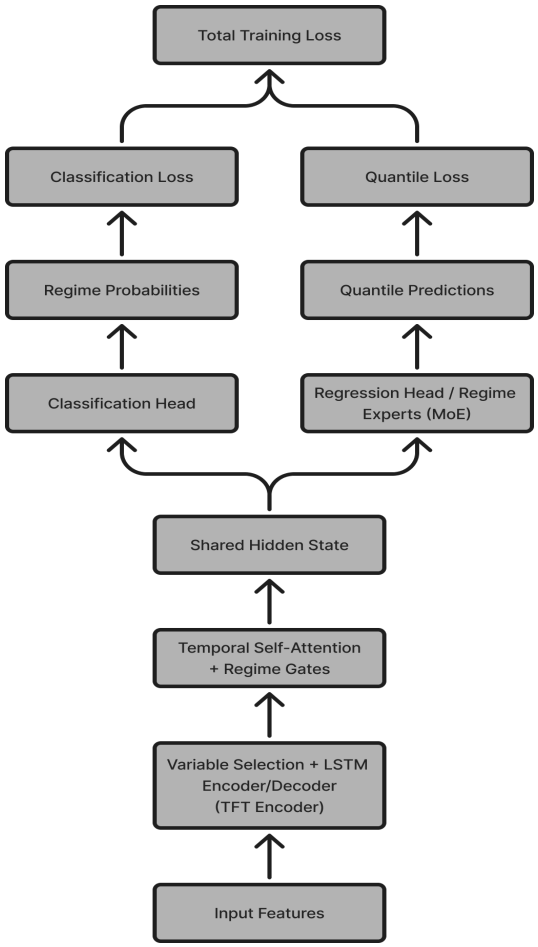

Figure 2. Modified TFT architecture. The base TFT encoder (Variable Selection + LSTM) feeds into temporal self-attention augmented with regime-conditional gates. The shared hidden state branches to an optional diagnostic classification head (regime prediction) and a regression head with an optional mixture-of-experts. When enabled, both losses contribute to the total training objective.

Regime-Aware Attention. We modify the interpretable multi-head attention mechanism to explicitly condition on the market regime. Each attention head is scaled by a regime-specific learned gate:

$$\tilde{a}_h = \sigma(g_{r,h}) \cdot a_h \tag{3}$$

where $g_{r,h}$ is the learned gate regime r and head h. This allows the heads to specialize, for example, by one head being amplified in high-volatility periods, and another during times of low-volatility. Regime assignment uses VIX thresholds for interpretability. This adds minimal parame-

ters (4 total for 2 regimes with 2 attention heads) while providing a direct mechanism for regime-dependent attention behavior.

Staleness-Aware Attention. We modify attention scores to explicitly penalize stale timesteps before the softmax operation:

$$\tilde{a}_{ij} = a_{ij} - \lambda \cdot d(s_j) \quad (4)$$

where a_{ij} is the raw attention logit from query i to key j, λ is a learned staleness weight, s_j is days since the most recent macro update at timestep j, and $d(.)$ is a decay function. We use log-normalized decay where staleness values are scaled to [0,1] across the encoder window. This encourages the model to discount information from timesteps where macro data has not been updated. We additionally test staleness as input features: a continuous counter (day since update) and a sparse binary flag (1 on release days only).

3.4. Evaluation Framework

We evaluate models using quantile loss, directional accuracy, and Sharpe ratio of a long-only strategy, alongside standard regression metrics (RMSE, MAE) and financial performance measures (hit rate, maximum drawdown).

Beyond aggregate metrics, we monitor training dynamics by tracking gradient flow through each layer, prediction variance, and directional bias over epochs. We classify evaluation windows by prediction quality (healthy vs. degraded/collapsed) based on variance, directional bias, and correlation with actual returns.

For select configurations, we employ rolling evaluation with 10-year training, 1-year validation, and 1-year test windows, stepping forward annually. This produces performance distributions across different market conditions rather than single-point estimates from one fixed split.

4. Results

4.1. Baseline Comparison

Table 1 presents the performance of ARIMAX and LSTM baselines alongside TFT variants on directional accuracy and Sharpe ratio metrics. ARIMAX achieves 54.0% directional accuracy (+0.4% excess) with a Sharpe ratio of 1.56. LSTM achieves 53.3% directional accuracy (-0.3% excess) with a Sharpe ratio of 0.39. TFT daily single-step prediction improves to 55.2% (+1.6% excess, Sharpe 1.21) with proper checkpoint selection. Multi-horizon daily prediction (10-step average) achieves 56.1% (+2.1% excess, Sharpe 2.08). Weekly TFT reaches 58.1% (+2.1% excess, Sharpe 1.05) and regime-aware attention extends this to 59.4% (+3.4% excess, Sharpe 1.22) on fixed-split evaluation, though this improvement does not generalize to rolling evaluation.

Model	Dir. Acc.	Baseline	Excess	Sharpe
ARIMAX(3,0,3)	54.0%	53.6%	+0.4%	1.56
LSTM	53.3%	53.6%	-0.3%	0.39
TFT (daily)	55.2%	53.6%	+1.6%	1.21
TFT (daily, 10-step)	56.1%	53.6%	+2.1%	2.08
TFT (weekly)	58.1%	56.0%	+2.1%	1.05
TFT + Regime Attn	59.4%	56.0%	+3.4%	1.22

Table 1. Performance on fixed-split test set (2020-2025). Baseline is frequency-matched naive prediction rate (53.6% for daily, 56.0% for weekly). TFT checkpoints selected by validation prediction diversity.

4.2. Prediction Collapse and Gradient Analysis

Preliminary testing revealed that TFT models with hidden dimensions above 18 consistently produce degenerate predictions, outputting near-constant positive values regardless of market conditions. We term this failure mode "prediction collapse", characterized by low output variance and persistent directional bias toward positive returns. Constraining hidden size to 16 or below mitigates but does not eliminate collapse.

Gradient analysis shows that there is an encoder-output disconnect (Figure 3). Output layer gradient norms collapse within the first 10 epochs and remain flat, while LSTM encoder and decoder gradients continue increasing throughout training. This pattern, consistent across experiments, indicates the encoder learns meaningful representations that the output layer fails to translate into predictions.

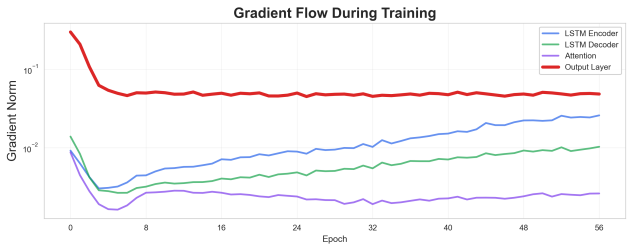

Figure 3. Gradient norm by layer during training. The output layer (red) collapses within the first 10 epochs and remains flat, while LSTM encoder and decoder gradients (blue, green) continue increasing throughout training. This pattern, consistent across experiments, indicates the encoder learns meaningful representations that the output layer fails to translate into predictions.

That gap contributes to why validation loss is a poor proxy for model quality. Quantile loss is anti-correlated with directional accuracy (r = -0.46) and prediction diversity (r = -0.64), meaning models achieve the lowest loss by collapsing to constant predictions that exploit market drift. Validation loss converged tightly (0.39-0.40) across experiments regardless of downstream performance. We therefore evaluate checkpoints by directional accuracy and prediction

standard deviation rather than validation loss.

4.3. Attention and Feature Selection Analysis

Attention mechanisms detect regime shifts at both frequencies despite prediction collapse. At daily frequency, 23 or 30 baseline experiments show attention pattern shifts at the 2022 to 2023 transition, corresponding to the Federal Reserve policy pivot. Weekly models show similar behavior with 12 of 24 experiments detecting shifts at the same transition. These shifts manifest as changes in temporal attention distribution, with volatile periods (2020, 2022, 2025) exhibiting increased recency bias compared to stable periods (Figure 4).

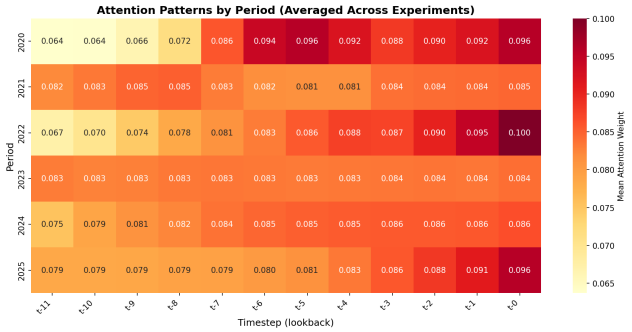

Figure 4. Temporal attention patterns across market regimes (weekly models, averaged across 24 experiments). Volatile periods (2020, 2022, 2025) exhibit increased recency bias, demonstrating the model's ability to adapt attention focus to market conditions.

The Variable Selection Network adapts feature weights appropriately across regimes (Figure 5). VIX weight peaks at 0.56 during the 2022 bear market before declining to 0.33 in 2024, while yield spread rises sharply to 0.37 in 2024 following rate hikes. Daily models show similar adaptive patterns, though with weaker yield spread emphasis (mean weight 0.12 versus 0.22 for weekly) and higher inflation weight (0.15 versus 0.05). Daily heatmaps are shown in Appendix 6

However, these attention metrics predict performance only at weekly frequency. For daily models, correlations between attention metrics and outcomes are weak: entropy trend versus directional accuracy (r = 0.05), VSN concentration variability versus composite score (r = 0.26). For weekly models, correlations are substantial: entropy trend versus directional accuracy (r = 0.60), entropy trend versus collapse rate (r = -0.69), and VSN concentration variability versus composite score (r = 0.80). See Appendix 6 for full correlation matrices.

This frequency dependence reinforces our temporal aggregation finding. Daily noise obscures the relationship between attention behavior and prediction quality. At

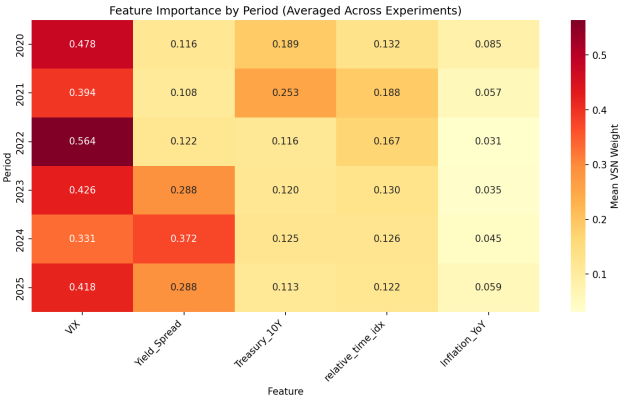

Figure 5. Variable Selection Network feature weights by period (weekly models, averaged across 24 experiments). VIX weight peaks during the 2022 bear market (0.56), while yield spread rises sharply in 2024 (0.37) following rate hikes, demonstrating adaptive feature selection across market regimes.

weekly frequency, models that adaptively shift attention focus achieve better performance and less collapse.

4.4. Architectural Modifications

Loss Function Penalties. Directional diversity penalties rescue suboptimal configurations but hurt those that are already strong. Adding dirw=1.0 to a weak baseline (h=16, dropout=0.10) improved the Sharpe ratio by 45% (0.72 to 1.05). However, applying the same penalty to an already-optimized configuration (h=16, dropout=0.25) degraded Sharpe by 28%. This configuration dependence limits the penalty's practical utility.

Regime-Conditional Output. We tested 48 mixture-of-experts configurations with parallel expert heads per regime. All configurations exhibited expert collapse: experts converged to static per-regime biases rather than learning sample-specific predictions. The router successfully learned regime signal (VIX correlation r = 0.83-0.87), confirming the hidden state contains regime information, but experts could not extract predictions from it.

Classification Head Diagnostic. To isolate whether collapse stems from the regression objective, we added a parallel classification head reading from the same encoder. The head achieved 100% accuracy on VIX-based regime classification but only base-rate accuracy (54%) on direction prediction. This confirms the encoder learns meaningful representations of the market state, but these representations do not contain a directional signal.

Regime-Aware Attention. Per-head attention gating conditioned on VIX regime learns interpretable behavior: high-volatility gates amplify attention (0.57), low-volatility gates dampen it (0.46) (Figure 6). The gates reshape temporal attention patterns, reducing recency bias during crises

and producing more uniform attention in stable markets (Figure 14. This gating induces dramatic VSN adaptation, with VIX weight reaching 0.74 during bear markets versus 0.31 at baseline, demonstrating that regime gates affect the entire model's feature selection, not just the attention layer (Figure 13). Fixed-split evaluation showed +1.5% directional accuracy, but rolling evaluation showed no improvement, suggesting fixed-split gains do not generalize.

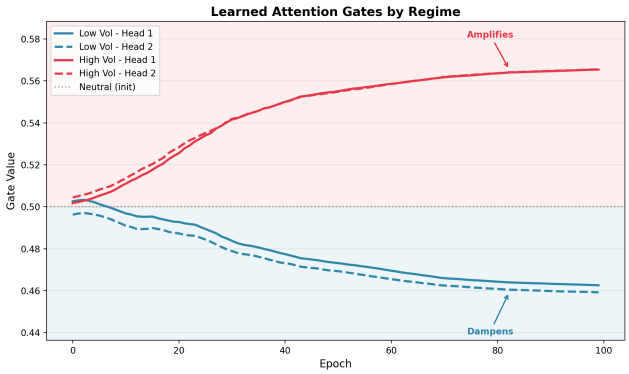

Figure 6. Learned attention gate values by regime during training. High-volatility gates (red) converge to amplify attention (0.56-0.57), while low-volatility gates (blue) dampen it (0.46), starting from neutral initialization (0.5). Both attention heads learn consistent regime-specific behavior.

Staleness-Aware Attention. Continuous staleness features (days since macro update) cause collapse regardless of whether encoded as inputs or attention penalties. Attention entropy drops from 2.48 to 1.37 as models exploit the recency gradient inherent in forward-filled data. Sparse binary release flags avoid collapse, but provide no predictive value, matching baseline performance.

Combinations. Combining modifications generally degraded performance. Regime output with loss penalties fell below 50% directional accuracy. Directional penalties on weekly models caused complete unidirectional collapse. Staleness attention combined with regime attention showed inconsistent results: daily models overcorrected to 47% accuracy with mostly negative predictions, while weekly models achieved 56.7% accuracy, worse than regime attention alone (59.4%). These negative results suggest that modifications interact unpredictably rather than providing an additive benefit.

4.5. Temporal Aggregation

Temporal aggregation enables signal extraction from mixed-frequency financial data. Weekly frequency and multi-horizon daily prediction both achieve approximately 2% excess directional accuracy over naive baselines, while daily single-step prediction shows minimal excess.

On fixed-split evaluation (2020-2025 test), weekly models achieve 58.1% directional accuracy versus a 56.0% baseline (+2.1% excess). Daily models with proper checkpoint selection achieve 55.2% accuracy on single-step prediction (+1.6% excess). Extending to multi-horizon prediction (10-step) improves daily performance to 56.1% average (+2.1% excess), matching weekly results. Rolling evaluation across nine years (2016-2024) confirms weekly superiority: 59.1% $\pm$ 8.3% accuracy (+2.3% excess) versus daily 53.3% $\pm$ 5.2% (-0.6% excess with validation loss checkpoints). Both frequencies fail during the 2022 bear market (˜40% accuracy), revealing limitations when regime shifts violate training distribution assumptions.

The frequency dependence is asymmetric: longer prediction horizons improve daily models (1-step to 10-step: +0.5pp accuracy) but not weekly models (1-step optimal). This suggests daily noise is reducible through aggregation, while weekly data is already near the signal ceiling for this feature set.

These results suggest a noise floor in daily returns that single-step prediction cannot overcome with this feature set. Whether aggregation occurs through frequency (weekly) or multi-task learning (multi-horizon), approximately 2% excess accuracy represents the extractable signal from our mixed-frequency macroeconomic features on S&P 500 returns.

5. Discussion

The success of temporal aggregation reveals fundamental structure in mixed-frequency financial forecasting: daily equity returns contain a noise floor that single-step prediction cannot overcome, but this noise is reducible through either frequency transformation or multi-horizon averaging. Both approaches achieve similar excess accuracy, suggesting they access the same underlying signal. This establishes temporal aggregation as a first-order design decision in mixed-frequency forecasting, comparable in importance to model architecture.

The encoder-output gradient disconnect we observe suggests a limitation in how transformers handle extremely noisy sequence-to-point regression. Attention mechanisms successfully detect regime transitions and adapt feature selection, demonstrating that encoder representations contain meaningful structure. However, output layer gradients dominate attention gradients by 50-100x during training, preventing the model from translating detected patterns into stable predictions. This disconnect explains why regime-conditional output layers achieve perfect regime classification yet fail at directional prediction despite reading from identical hidden states. The anti-correlation between quantile loss and directional accuracy compounds this issue, as the standard training objective actively selects for collapsed models.

Our negative results on architectural modifications high-

light the difficulty of improving transformer performance through model design alone. Staleness-aware attention degrades performance because forward-filled data creates spurious temporal gradients that attention mechanisms exploit. Regime-conditional outputs fail because the regression objective provides insufficient supervision for expert specialization. Loss penalties show configuration-dependent effects because they address symptoms rather than root causes. These failures suggest that transformers need domain-specific preprocessing and training strategies, not just architectural tweaks, when applied to financial forecasting.

Future work should explore whether hierarchical encoder architectures with separate pathways for daily and monthly data can eliminate forward-fill artifacts, and whether alternative aggregation strategies beyond simple weekly averaging or multi-horizon prediction can further improve performance.

6. Conclusion

We evaluate Temporal Fusion Transformers for S&P 500 return forecasting using mixed-frequency macroeconomic data. Temporal aggregation through weekly frequency or multi-horizon daily prediction achieves approximately 2% excess directional accuracy. Our diagnostic analysis reveals that attention mechanisms successfully detect regime transitions despite an encoder-output gradient disconnect that prevents these patterns from influencing predictions. This disconnect explains failures in regime-conditional architectures and necessitates checkpoint selection by prediction diversity rather than validation loss. Architectural modifications targeting staleness and regime adaptation show limited benefit. These findings establish practical guidance for applying transformers to financial forecasting while documenting fundamental limitations in extremely noisy regression domains.

References

[1] E. F. Fama, "Efficient capital markets: A review of theory and empirical work," *Journal of Finance*, vol. 25, no. 2, pp. 383–417, 1970. 1

[2] P. Chen and C.-H. Shen, "Regime-switching models: Capturing structural changes in time series," in *Proc. SAS Global Forum*, 2018. 1

[3] A. Ang and A. Timmermann, "Regime changes and financial markets," *Annual Review of Financial Economics*, vol. 4, no. 1, pp. 313–337, 2012. 1

[4] B. Lim, S. Arik, N. Loeff, and T. Pfister, "Temporal fusion transformers for interpretable multi-horizon time series forecasting," *International Journal of Forecasting*, vol. 37, no. 4, pp. 1748–1764, 2021. 1, 2, 3

[5] H. Wu, J. Xu, J. Wang, and M. Long, "Autoformer: Decomposition transformers with auto-correlation for long-term series forecasting," in *Advances in Neural Information Processing Systems* (M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, and J. W. Vaughan, eds.), vol. 34, pp. 22419–22430, Curran Associates, Inc., 2021. 2, 3

[6] H. Zhou, S. Zhang, J. Peng, S. Zhang, J. Li, H. Xiong, and W. Zhang, "Informer: Beyond efficient transformer for long sequence time-series forecasting," *CoRR*, vol. abs/2012.07436, 2020. 2

[7] P. Liu, B. Wu, Y. Hu, N. Li, T. Dai, J. Bao, and S. Xia, "Timebridge: Non-stationarity matters for long-term time series forecasting," in *Proc. International Conference on Machine Learning (ICML)*, 2025. 2

[8] E. Ghysels, P. Santa-Clara, and R. Valkanov, "The midas touch: Mixed data sampling regression models," *Journal of Financial Economics*, vol. 78, no. 2, pp. 213–244, 2005. 2, 3

[9] T. Fischer and C. Krauss, "Deep learning with long short-term memory networks for financial market predictions," *European Journal of Operational Research*, vol. 270, no. 2, pp. 654–669, 2018. 2

[10] International Finance Corporation, "Leveraging machine learning to enhance credit data quality," technical report, World Bank Group, 2025. 3

[11] "Temporal fusion transformers for enhanced multivariate time series forecasting in stock market prediction," *International Journal of Advanced Computer Science and Applications*, vol. 15, no. 7, pp. 1–8, 2024. TFT obtained a remarkable SMAPE of 0.0022. 3

[12] Y. Chen *et al.*, "Economic system forecasting based on temporal fusion transformers: Multi-dimensional evaluation and cross-model comparative analysis," *Neurocomputing*, vol. 549, p. 126433, 2023. 3

[13] Y. Pei, J. Cartlidge, A. Mandal, D. Gold, E. Marcilio, and R. Mazzon, "Cross-modal temporal fusion for financial market forecasting," in *Proc. European Conference on Artificial Intelligence (ECAI)*, (Bologna, Italy), 2025. Manuscript accepted to PAIS at ECAI-2025. 3

[14] F. R. B. of St. Louis, "ALFRED: Archival Federal Reserve Economic Data." https://alfred.stlouisfed.org/, 2025. 3

[15] G. E. P. Box and G. C. Tiao, "Intervention analysis with applications to economic and environmental

problems," *Journal of the American Statistical Association*, vol. 70, no. 349, pp. 70–79, 1975. 3

[16] J. Beitner and contributors, "PyTorch Forecasting: QuantileLoss_pkg Documentation." https://pytorch-forecasting.readthedocs.io/en/stable/api/pytorch_forecasting.metrics._quantile_pkg._quantile_loss_pkg.QuantileLoss_pkg.html, 2025. 4

[17] N. Shazeer, A. Mirhoseini, K. Maziarz, A. Davis, Q. V. Le, G. E. Hinton, and J. Dean, "Outrageously large neural networks: The sparsely-gated mixture-of-experts layer," *CoRR*, vol. abs/1701.06538, 2017. 4

[18] R. Caruana, "Multitask learning," *Machine Learning*, vol. 28, pp. 41–75, 1997. 4

A. Attention Pattern Analysis

This appendix provides additional visualizations supporting the attention mechanism analysis in Section 4.

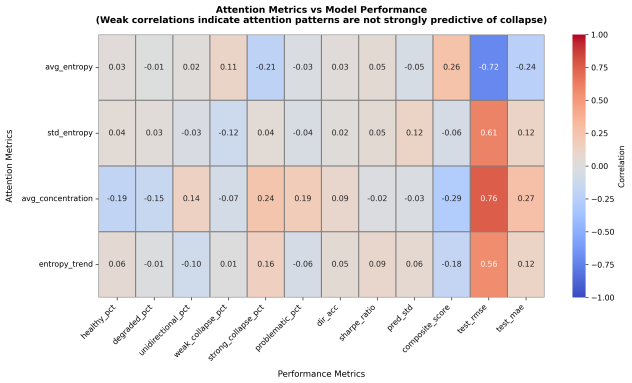

Figure 7. Temporal attention metrics versus model performance (daily models, 30 experiments). Weak correlations indicate attention patterns do not predict outcomes at a daily frequency.

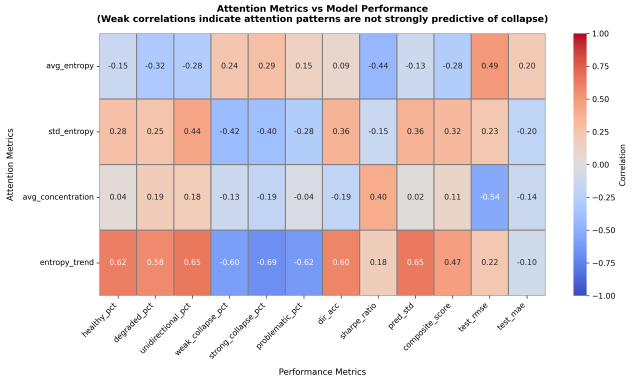

Figure 8. Temporal attention metrics versus model performance (weekly models, 24 experiments). Strong correlations (entropy_trend ↔ dir_acc = 0.60) indicate attention dynamics predict outcomes at weekly frequency.

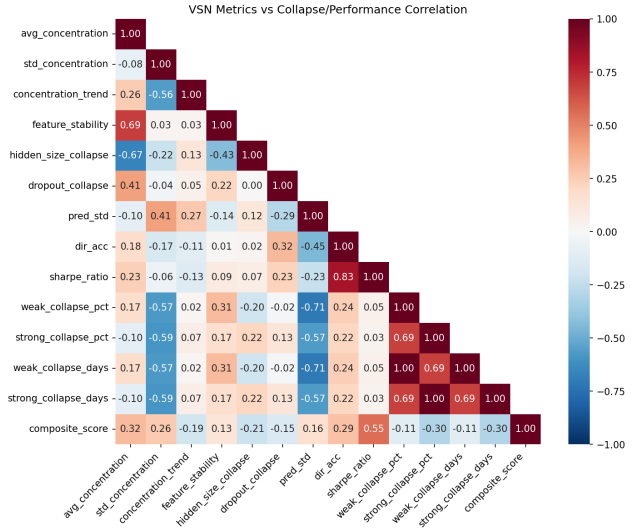

Figure 9. Variable Selection Network metrics versus model performance (daily models, 30 experiments). Weak correlations indicate feature selection dynamics do not strongly predict outcomes at a daily frequency.

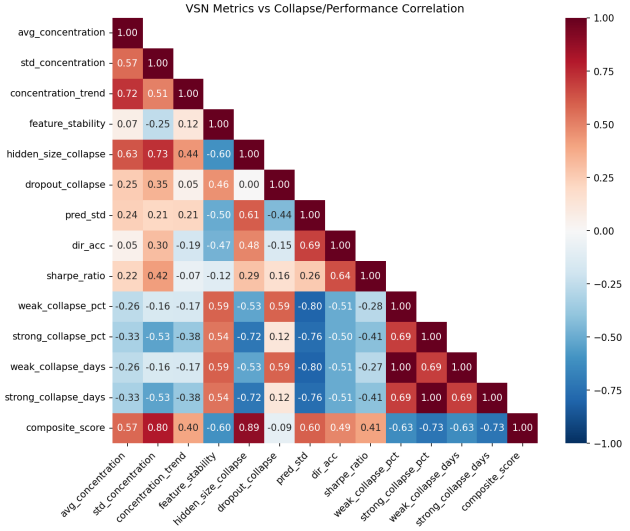

Figure 10. Variable Selection Network metrics versus model performance (weekly models, 24 experiments). Strong correlations (std_concentration $\leftrightarrow$ composite_score = 0.80) indicate feature selection variability predicts outcomes at weekly frequency.

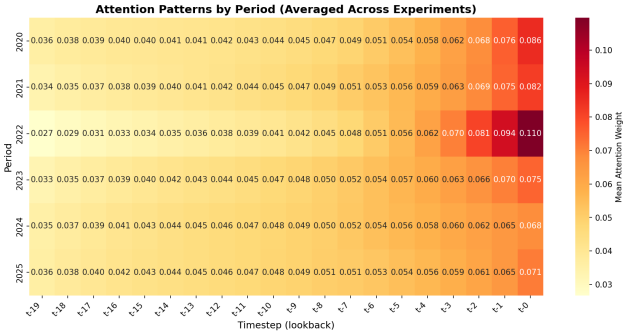

Figure 11. Temporal attention patterns across market regimes (daily models, averaged across 30 experiments). Shows similar regime detection to weekly models but with weaker correlation to performance outcomes.

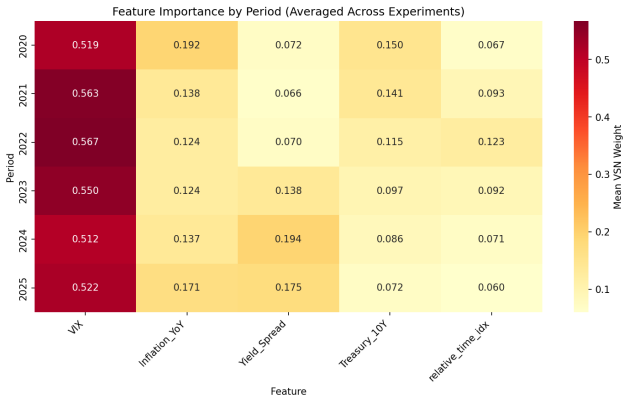

Figure 12. Variable Selection Network feature weights by period (daily models, averaged across 30 experiments). VIX dominates across all periods with more stable weights compared to weekly models.

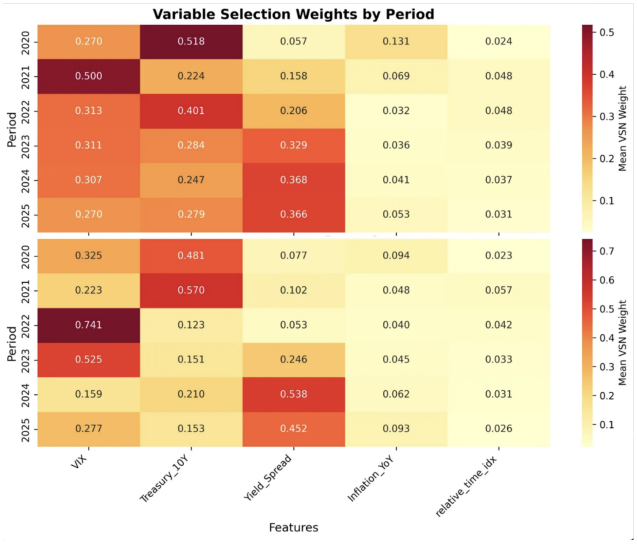

Figure 13. Variable Selection Network weights with and without regime-aware attention (single weekly configuration). Regime attention induces dramatic adaptation: VIX weight reaches 0.74 during the 2022 bear market versus 0.31 baseline, and yield spread dominates 2024 (0.54 versus 0.37 baseline).

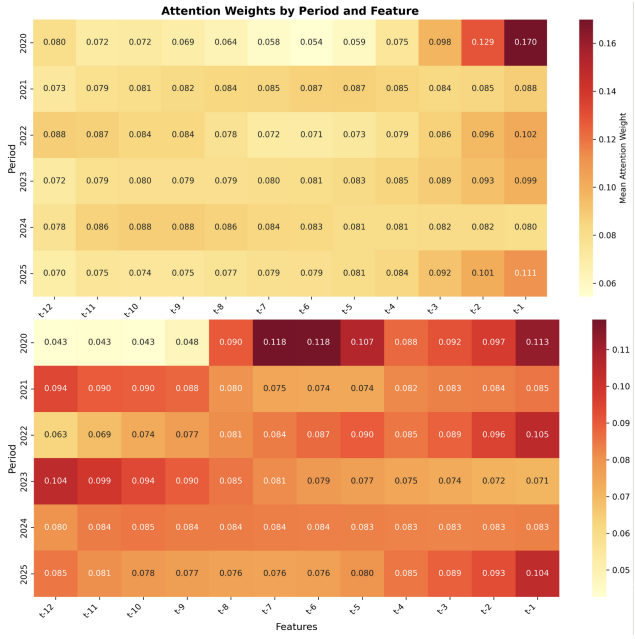

Figure 14. Temporal attention patterns with and without regime-aware attention (single weekly configuration). Regime gates reshape attention strategies: reducing recency bias during crisis periods (2020) and producing more uniform attention in stable markets (2024).