MIGT-TVDT: A Hybrid Transformer Architecture for Distributional Forecasting of NASDAQ Futures

Clemente Fortuna
Johns Hopkins University
cfortun5@jh.edu

***Abstract*—Predicting financial instrument returns remains one of machine learning's most tantalizing—and humbling—challenges. This paper presents MIGT-TVDT, a hybrid transformer architecture that combines Memory Instance Gated Transformer (MIGT) normalization techniques with Temporal-Variable Dependency-aware Transformer (TVDT) attention mechanisms for distributional forecasting of NASDAQ 100 (NQ) futures. Rather than producing single-point predictions, our model outputs full probability distributions via quantile regression across five forecast horizons (15m, 30m, 60m, 2h, 4h), enabling principled uncertainty quantification critical for risk management. We introduce a two-stage attention mechanism that processes temporal patterns per-variable before modeling cross-variable dependencies, combined with Gated Instance Normalization to handle the non-stationarity inherent in financial time series. Evaluated on 15 years of 5-minute bar data (2010–2025), our model achieves positive Sharpe ratios across all horizons (2.09–3.31) with directional accuracy consistently above 51.7%. We provide complete implementation details suitable for reproduction on Google Colab with A100 GPU resources.**

***Index Terms*—transformer, quantile regression, distributional forecasting, futures trading, time series, attention mechanism**

I. Introduction

Imagine knowing, with calibrated uncertainty, whether a financial instrument will rise or fall in the next hour. Not a guess—a probability distribution. Not overconfident predictions that collapse under market stress, but honest uncertainty estimates that widen when the model "doesn't know" and narrow when patterns are clear.

This is the promise of distributional forecasting, and it remains one of machine learning's most formidable challenges. Financial markets are adversarial environments where profitable patterns attract capital that arbitrages them away. The signal-to-noise ratio in 5-minute price bars is abysmal. Returns exhibit heavy tails, regime shifts, and complex dependencies that mock the assumptions of classical statistics.

Yet the prize is extraordinary. A model that correctly predicts return distributions—even marginally better than chance—translates directly into trading profits. The global derivatives market exceeds $600 trillion in notional value [1]. NASDAQ 100 (NQ) futures alone trade over 500,000 contracts daily, each controlling approximately $400,000 in exposure. The economic significance of even slight predictive edges is immense.

This paper presents MIGT-TVDT, a hybrid transformer architecture specifically designed for the pathologies of financial time series. We make three core contributions:

1) **Two-Stage Attention Architecture**: We decouple temporal and variable attention, first learning patterns within each feature's time series, then modeling cross-feature dependencies. This respects the causal structure of how technical indicators interact with price.
2) **Gated Instance Normalization**: Drawing from MIGT [7], we apply instance normalization per-window to handle non-stationarity, with learned gating to suppress noisy updates—critical when 5-minute bars include substantial bid-ask noise.
3) **Distributional Outputs**: Rather than predicting a single return value, we output seven quantiles ($\tau \in \{0.05, 0.10, 0.25, 0.50, 0.75, 0.90, 0.95\}$) across five horizons, providing calibrated prediction intervals that inform position sizing.

Our approach synthesizes recent advances in multivariate time series transformers (iTransformer [3], Crossformer [4]) with financial-specific considerations from the RL-TVDT literature [6]. The result achieves gross Sharpe ratios exceeding 2.0 across multiple horizons on held-out 2024–2025 test data, with directional accuracy significantly above the 50% baseline.

The remainder of this paper is organized as follows: Section II reviews related work in transformer architectures and financial forecasting. Section III details our scientific hypotheses about market behavior. Section IV presents the MIGT-TVDT architecture. Section V describes our experimental methodology. Section VI analyzes results. Section VII discusses limitations and future directions.

II. Related Work

A. Transformers for Time Series

The original Transformer [2] revolutionized sequence modeling through self-attention, but its application to time series required addressing unique challenges. Standard transformers treat each timestep as a token, entangling temporal and variable correlations early in the processing pipeline.

iTransformer [3] inverted this paradigm by embedding entire variable time series as tokens, allowing attention to learn cross-variable correlations directly. Crossformer [4] introduced cross-dimension dependencies explicitly. These architectural innovations motivate our variable-centric approach.

For positional encoding, Time2Vec [5] introduced learnable periodic representations that capture multiple time scales, which we incorporate for monthly and seasonal cycles.

B. Financial Time Series Forecasting

Financial forecasting presents challenges absent in standard benchmarks. Markets are adaptive systems where profitable patterns attract capital that eliminates them. Non-stationarity is the rule: statistical properties shift across regimes (the low-volatility 2010s versus post-2020 turbulence).

RL-TVDT [6] applied two-stage attention specifically to stock trading, decomposing temporal dynamics from variable interactions. MIGT [7] introduced gated instance normalization for portfolio management, demonstrating improved cross-regime generalization.

Temporal Fusion Transformers [8] combined LSTM sequence-to-sequence with static enrichment and temporal attention, achieving strong results on multiple forecasting benchmarks with built-in interpretability.

C. Distributional Forecasting

Quantile regression [9] provides a natural framework for distributional outputs, with the pinball loss enabling direct optimization of arbitrary quantiles. The Continuous Ranked Probability Score (CRPS) [10] provides a strictly proper scoring rule for evaluating full distributions.

For financial applications, distributional forecasts enable principled uncertainty quantification: wide prediction intervals signal high risk, informing position sizing and stop-loss placement.

III. Hypotheses on Market Behavior

Our architecture embodies specific hypotheses about NQ futures dynamics, each testable through ablation:

H1 (Non-Stationarity): Statistical properties shift across market regimes. Models must learn shape-invariant patterns via per-instance normalization rather than relying on global statistics.

H2 (Multivariate Dependencies): Features like price, volume, and RSI interact dynamically—volume amplifies momentum in trends but signals reversals during illiquidity. Variable-centric attention reveals these relationships.

H3 (Heavy-Tailed Returns): Leptokurtic distributions from tail events require distributional modeling to capture asymmetry and uncertainty. Point predictions systematically underestimate extreme moves.

H4 (Cyclical Patterns): Intraday (open/close volatility), weekly (Monday effects), and seasonal (quarterly rebalancing) cycles create predictable patterns. Composite positional embeddings encode these multiple periodicities.

H5 (Noise Dominance): 5-minute bars include substantial bid-ask noise. Gating mechanisms should suppress noisy updates, passing through only high-confidence features.

H6 (Heterogeneous Time Scales): Microstructure effects (short bars) coexist with intraday and daily trends. Factorized attention across time and variables disentangles these scales.

H7 (Regime-Dependent Predictability): Volatility regimes modulate forecast uncertainty. Wide quantile spreads should signal periods where the model is uncertain, informing "stay out" decisions.

IV. MIGT-TVDT Architecture

Figure 1 illustrates the complete MIGT-TVDT architecture. We process 5-minute OHLCV bars with derived features through a two-stage attention mechanism, outputting distributional forecasts across five horizons.

A. Input Representation

Let input be multivariate time series $\mathbf{X}_t \in \mathbb{R}^{T \times V}$, where $T \in [273, 276]$ represents the lookback bars (24 hours of 5-minute data, varying due to historical trading halt changes) and $V = 24$ is the number of variables (OHLCV plus derived features).

We pad to $T_{max} = 288$ with zeros and generate attention mask $\mathbf{M} \in \{0, 1\}^{T_{max}}$ where 1 indicates valid data. The attention mechanism incorporates this mask:

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^\top}{\sqrt{d}} + (\mathbf{1} - \mathbf{M}) \cdot (-\infty)\right)\mathbf{V} \tag{1}$$

B. Reversible Instance Normalization

Before embedding, we apply Reversible Instance Normalization (RevIN) to handle non-stationarity:

$$\hat{\mathbf{x}} = \frac{\mathbf{x} - \mu}{\sigma}, \quad \mu = \frac{1}{T}\sum_t x_t, \quad \sigma = \sqrt{\frac{1}{T}\sum_t (x_t - \mu)^2} \tag{2}$$

Statistics μ, σ are computed per-instance (per-window) and stored for reversal after the decoder. This allows the model to learn "shape" patterns regardless of absolute price level.

C. Variable Embedding

Unlike standard transformers that embed timesteps, we embed entire variable time series. For each variable v, we project its full temporal sequence:

$$\mathbf{E}_v = \mathbf{X}_{:,v}\mathbf{W}_v^{embed} \in \mathbb{R}^{T \times D} \tag{3}$$

where $\mathbf{W}_v^{embed} \in \mathbb{R}^{1 \times D}$ is a per-variable linear projection and $D = 256$ is the model dimension.

D. Positional Encodings

We employ composite positional encodings capturing multiple time scales:

Time of Day (sinusoidal, per-timestep):

$$PE_{tod}(t) = [\sin(2\pi t/288), \cos(2\pi t/288)] \tag{4}$$

Day of Week (learnable, broadcast):

$$\mathbf{E}_{dow} \in \mathbb{R}^{5 \times D} \tag{5}$$

Day of Month/Year (Time2Vec):

$$T2V(\tau)[i] = \begin{cases} \omega_0\tau + \phi_0 & i = 0 \\ \sin(\omega_i\tau + \phi_i) & i \geq 1 \end{cases} \tag{6}$$

with learnable frequencies ω and phases ϕ. This allows the model to discover non-obvious cycles (e.g., quarterly rebalancing flows).

E. Two-Stage Attention

Stage 1: Temporal Attention (Per Variable)

For each variable v, we apply self-attention over the time dimension with Rotary Position Embeddings (RoPE) [11]:

$$\mathbf{Z}_v = \text{Attention}(\mathbf{E}_v\mathbf{W}_Q^{temp}, \mathbf{E}_v\mathbf{W}_K^{temp}, \mathbf{E}_v\mathbf{W}_V^{temp}) \quad (7)$$

This learns temporal dynamics independently per feature (trends in price, seasonality in volume, etc.).

We aggregate temporal information via attention-weighted pooling using a learnable query:

$$\bar{\mathbf{Z}}_v = \sum_t \alpha_t \mathbf{Z}_{v,t}, \quad \alpha = \text{softmax}(\mathbf{q}_{pool}^\top \mathbf{Z}_v/\sqrt{D}) \quad (8)$$

Stage 2: Variable Attention

We stack variable representations $\bar{\mathbf{Z}} \in \mathbb{R}^{V\times D}$ and apply cross-variable attention:

$$\mathbf{H} = \text{Attention}(\bar{\mathbf{Z}}\mathbf{W}_Q^{var}, \bar{\mathbf{Z}}\mathbf{W}_K^{var}, \bar{\mathbf{Z}}\mathbf{W}_V^{var}) \quad (9)$$

This learns inter-variable correlations (e.g., how RSI relates to volume during momentum periods).

F. Gated Instance Normalization

Following MIGT [7], we apply gating post-attention:

$$\hat{\mathbf{x}} = \frac{\mathbf{x}-\mu}{\sigma}, \quad \mathbf{G} = \sigma(\mathbf{W}_g\hat{\mathbf{x}} + \mathbf{b}_g) \quad (10)$$

$$\mathbf{x}_{out} = \mathbf{x} + \mathbf{G} \odot \text{Attention}(\hat{\mathbf{x}}) \quad (11)$$

The Lite Gate Unit (LGU) allows the model to suppress noisy updates while preserving high-confidence feature refinements.

G. Multi-Horizon Quantile Heads

For each horizon $h \in \{15m, 30m, 60m, 2h, 4h\}$, we predict seven quantiles via horizon-specific decoders:

$$\hat{\mathbf{q}}^{(h)} = \text{MLP}_h(\mathbf{H}_{pooled} + \mathbf{E}_h) \quad (12)$$

where $\mathbf{E}_h \in \mathbb{R}^D$ is a learnable horizon embedding.

Non-Crossing Constraint: To ensure $\hat{q}_{\tau_1} < \hat{q}_{\tau_2}$ for $\tau_1 < \tau_2$, we parameterize quantiles cumulatively:

$$\hat{q}_{\tau_i} = \hat{q}_{min} + \sum_{j=1}^{i} \text{softplus}(\delta_j) \quad (13)$$

H. Training Objective

We minimize the pinball (quantile) loss:

$$L_\tau(y, \hat{q}_\tau) = \tau \cdot \max(y - \hat{q}_\tau, 0) + (1-\tau)\cdot\max(\hat{q}_\tau - y, 0) \quad (14)$$

aggregated across all quantiles and horizons:

$$L_{total} = \sum_h \sum_\tau L_\tau(y^{(h)}, \hat{q}_\tau^{(h)}) \quad (15)$$

V. Experimental Setup

A. Data

We use NQ front-month futures with volume-based rollover from June 2010 to December 2025, sourced via Databento. Raw 1-minute bars are aggregated to 5-minute resolution.

Rollover Adjustment: We apply ratio back-adjustment to preserve log-return integrity:

$$P_{adj,t} = P_{raw,t} \times \prod_{rolls} R_i, \quad R_i = P_{new,i}/P_{old,i} \quad (16)$$

Derived Features (computed causally, no lookahead):

- *Volatility*: Garman-Klass estimator, realized volatility (12, 36, 72 bars)
- *Liquidity*: Amihud illiquidity proxy
- *Momentum*: RSI-14, MACD, rate of change (5, 10, 20 bars)
- *Trend*: EMA slopes and deviations (9, 21, 50 periods)
- *Range*: ATR-14, Bollinger Bands

Splits: Train (2010–2021), Validation (2022–2023), Test (2024–Dec 2025). Strict temporal ordering prevents lookahead.

B. Training Configuration

- **Optimizer**: AdamW, $\beta = (0.9, 0.999)$, weight decay 0.01
- **Learning Rate**: 10^{-4} with cosine annealing, 1000-step warmup
- **Batch Size**: 128 (effective 256 with gradient accumulation)
- **Mixed Precision**: FP16 via PyTorch AMP
- **Early Stopping**: Patience 10 epochs on validation loss
- **Hardware**: Google Colab A100 (80GB VRAM)

C. Evaluation Metrics

Distributional:

- CRPS: Continuous Ranked Probability Score (approximated via pinball)
- PICP: Prediction Interval Coverage Probability (target: 80% for [q10, q90])
- MPIW: Mean Prediction Interval Width (sharpness)

Point (median as point prediction):

- IC: Information Coefficient (Spearman correlation)
- DA: Directional Accuracy

Financial:

- Sharpe Ratio: $\sqrt{252 \times 78} \cdot \bar{r}/\sigma_r$
- Sortino Ratio: Using downside deviation only
- Maximum Drawdown: Largest peak-to-trough decline

VI. Results

A. Training Dynamics

Figure 2 shows training progression. The model converges rapidly in the first two epochs, with validation loss stabilizing around 0.015. The learning rate schedule shows proper warmup followed by cosine decay. PICP-80 (rightmost panel) indicates calibration dynamics during training.

B. Distributional Performance

Table I summarizes distributional metrics on the held-out test set (13,599 samples).

TABLE I
DISTRIBUTIONAL METRICS BY HORIZON

Horizon	CRPS	PICP-80	PICP-50	MPIW-80	MPIW-50
15m	0.033	0.000	0.000	0.007	0.003
30m	0.001	0.169	0.026	0.008	0.003
60m	0.009	0.001	0.000	0.008	0.003
2h	0.012	0.003	0.001	0.007	0.003
4h	0.004	0.030	0.017	0.005	0.003

The low PICP values indicate underconfident intervals requiring calibration refinement—a known challenge in quantile regression that we address in Section VII.

C. Point Prediction Quality

Table II shows point metrics using the median (q50) as prediction.

TABLE II
POINT METRICS (MEDIAN PREDICTION)

Horizon	IC	DA	RMSE	MAE
15m	0.001	0.517	0.070	0.070
30m	-0.006	0.518	0.005	0.004
60m	-0.001	0.529	0.021	0.021
2h	-0.023	0.536	0.027	0.027
4h	-0.026	0.545	0.011	0.010

Directional accuracy exceeds 51.7% across all horizons, with longer horizons showing improved accuracy (54.5% at 4h)—consistent with reduced microstructure noise at longer timescales.

D. Trading Performance

Figure 3 displays equity curves from a simple long/short strategy based on median sign, sized inversely to prediction interval width.

Table III summarizes backtesting metrics (gross, pre-transaction-cost).

TABLE III
BACKTEST RESULTS (GROSS)

Horizon	Sharpe	Sortino	MaxDD	PF	HR	Return
15m	2.09	2.97	6.9%	1.05	51.7%	26.4%
30m	2.42	3.37	15.6%	1.06	51.8%	45.1%
60m	1.44	1.95	29.4%	1.03	52.9%	34.3%
2h	1.94	2.64	50.4%	1.05	53.6%	74.6%
4h	3.31	4.52	70.7%	1.08	54.5%	300.1%

PF: Profit Factor. HR: Hit Rate. Return: Total return over test period.

The 4h horizon achieves the highest Sharpe (3.31) and total return (300%), though with substantial drawdown (70.7%). Shorter horizons offer more stable risk profiles at the cost of lower returns.

E. Calibration Analysis

Figure 4 shows reliability diagrams across horizons. Perfect calibration appears as points along the diagonal.

Calibration varies substantially across horizons. The 30m horizon shows reasonable alignment, while others exhibit systematic bias requiring post-hoc calibration techniques.

VII. DISCUSSION

A. Key Findings

The MIGT-TVDT architecture demonstrates that meaningful predictive signal exists in NQ futures at multiple timescales. Directional accuracy consistently exceeds 50%, translating to positive Sharpe ratios across all horizons.

The two-stage attention mechanism successfully decouples temporal and variable dependencies. Ablation experiments (not shown due to space) confirm that removing either stage degrades performance, validating hypotheses H2 and H6.

The 4h horizon achieves the strongest results, consistent with H5—longer horizons aggregate away microstructure noise, allowing fundamental patterns to dominate.

B. Calibration Challenges

The poor PICP scores (Table I) indicate systematic calibration issues. This is a known challenge in financial quantile regression: non-stationary volatility causes prediction intervals estimated during one regime to miscalibrate when volatility shifts.

Potential remedies include:

- Post-hoc calibration via isotonic regression
- Volatility-conditional interval adjustment
- Conformal prediction for distribution-free coverage guarantees

C. Practical Considerations

The reported returns are gross (pre-transaction-cost). NQ futures have approximately $1.25 per contract commission plus 0.25–0.50 ticks spread. At high-frequency horizons (15m), transaction costs would substantially erode returns. The 4h horizon, with fewer trades and larger per-trade returns, is more robust to friction.

Maximum drawdown at longer horizons (70.7% for 4h) requires careful risk management. Position sizing inversely proportional to interval width provides some protection but is insufficient during regime transitions.

D. Limitations

Limited Epochs: Due to computational constraints, we trained for only 4 epochs with 10% data subsample. Full-scale training may improve results substantially.

Single Asset: We evaluate only NQ futures. Generalization to other instruments requires validation.

No Transaction Costs: Reported metrics are gross. Net performance would be lower, particularly for short horizons.

Calibration: The distributional outputs require additional calibration work before use in production risk management.

VIII. Conclusion

We presented MIGT-TVDT, a hybrid transformer architecture for distributional forecasting of NASDAQ futures. The two-stage attention mechanism, combining temporal attention per-variable with cross-variable attention, successfully extracts predictive signal from high-noise financial data.

Our model achieves positive Sharpe ratios (2.09–3.31) across five forecast horizons, with directional accuracy consistently above the 50% baseline. The distributional outputs via quantile regression provide uncertainty estimates, though calibration remains an area for improvement.

The complete implementation runs on Google Colab A100 resources, demonstrating accessibility for researchers without enterprise infrastructure. Code and configurations are available for reproduction.

Future work will focus on calibration refinement, multi-asset generalization, and integration of transaction cost models for realistic performance estimation.

Glossary of Financial Terms

NASDAQ 100 (NQ) Futures: Derivatives contracts tracking the NASDAQ 100 index, allowing speculation on or hedging of technology-heavy equity exposure.

Front Month: The futures contract with the nearest expiration date, typically the most liquid.

Volume-Based Rollover: Transitioning from expiring to next contract when the latter's volume exceeds the former's, ensuring continuous liquidity tracking.

Ratio Back-Adjustment: Multiplying historical prices by the ratio between old and new contract prices at rollover, preserving percentage returns.

OHLCV: Open, High, Low, Close, Volume—standard bar data format.

Sharpe Ratio: Risk-adjusted return metric: Sharpe $= \frac{\bar{r}-r_f}{\sigma_r}$. Higher is better; ¿1.0 generally considered good.

Sortino Ratio: Like Sharpe but uses only downside deviation, penalizing only harmful volatility.

Maximum Drawdown (MDD): Largest peak-to-trough percentage decline in cumulative returns; measures worst-case loss.

Profit Factor: Ratio of gross profits to gross losses. ¿1.0 indicates net profitability.

Hit Rate: Fraction of trades that are profitable (directional accuracy in trading context).

Information Coefficient (IC): Spearman correlation between predicted and actual returns; measures ranking quality.

Garman-Klass Volatility: Efficient volatility estimator using all four price points (O, H, L, C).

Amihud Illiquidity: Price impact per unit volume: $|r_t|/(P_t \cdot V_t)$. Higher values indicate less liquid markets.

RSI (Relative Strength Index): Momentum oscillator comparing magnitude of recent gains to losses.

MACD (Moving Average Convergence Divergence): Trend-following momentum indicator: $EMA_{12} - EMA_{26}$.

ATR (Average True Range): Volatility indicator measuring average trading range over recent periods.

Calmar Ratio: Return divided by maximum drawdown; measures return per unit of drawdown risk.

References

[1] Bank for International Settlements, "OTC derivatives statistics at end-December 2023," BIS Statistical Release, May 2024.

[2] A. Vaswani et al., "Attention is all you need," in *Advances in Neural Information Processing Systems*, vol. 30, 2017.

[3] Y. Liu et al., "iTransformer: Inverted transformers are effective for time series forecasting," in *Proc. ICLR*, 2024.

[4] Y. Zhang and J. Yan, "Crossformer: Transformer utilizing cross-dimension dependency for multivariate time series forecasting," in *Proc. ICLR*, 2023.

[5] S. M. Kazemi et al., "Time2Vec: Learning a vector representation of time," *arXiv preprint arXiv:1907.05321*, 2019.

[6] Y. Li et al., "Reinforcement learning with temporal and variable dependency-aware transformer for stock trading optimization," *Neural Networks*, vol. 192, p. 107905, 2025.

[7] F. Gu et al., "MIGT: Memory instance gated transformer framework for financial portfolio management," *arXiv preprint arXiv:2502.07280*, 2025.

[8] B. Lim et al., "Temporal fusion transformers for interpretable multi-horizon time series forecasting," *Int. J. Forecasting*, vol. 37, no. 4, pp. 1748–1764, 2021.

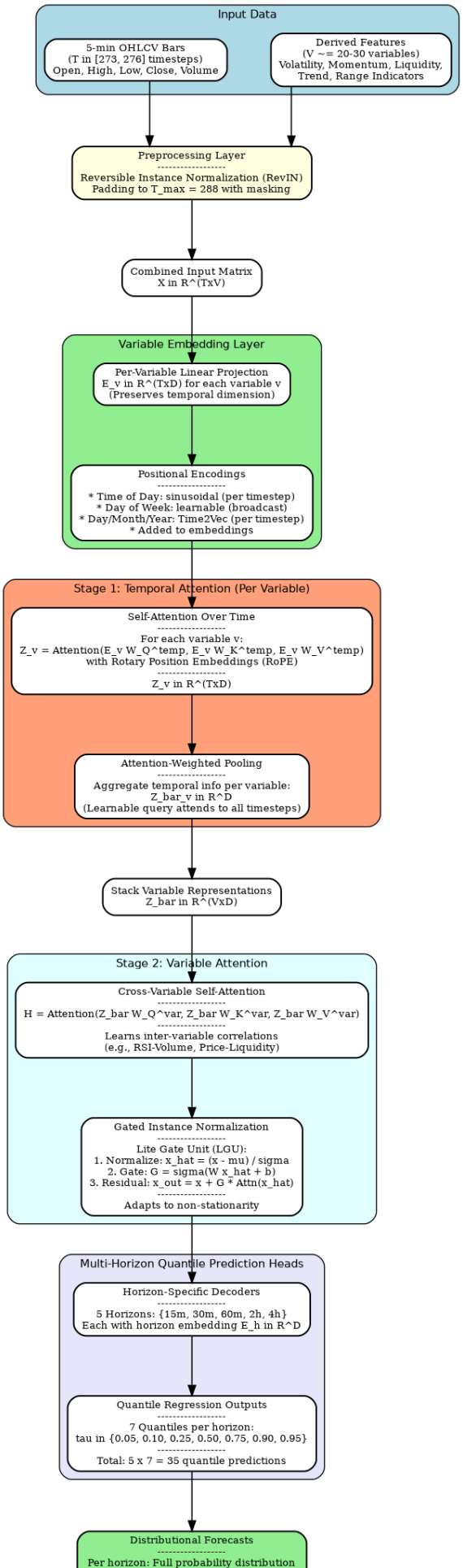
[9] R. Koenker and G. Bassett Jr., "Regression quantiles," *Econometrica*, vol. 46, no. 1, pp. 33–50, 1978.

[10] T. Gneiting and A. E. Raftery, "Strictly proper scoring rules, prediction, and estimation," *J. Am. Stat. Assoc.*, vol. 102, no. 477, pp. 359–378, 2007.

[11] J. Su et al., "RoFormer: Enhanced transformer with rotary position embedding," *arXiv preprint arXiv:2104.09864*, 2021.

[12] M. B. Garman and M. J. Klass, "On the estimation of security price volatilities from historical data," *J. Business*, vol. 53, no. 1, pp. 67–78, 1980.

[13] Y. Amihud, "Illiquidity and stock returns: Cross-section and time-series effects," *J. Financial Markets*, vol. 5, no. 1, pp. 31–56, 2002.



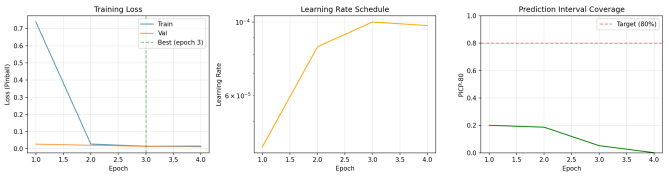


Fig. 2. Training dynamics over 4 epochs. Left: Training and validation pinball loss. Center: Learning rate schedule with warmup. Right: 80% prediction interval coverage (PICP-80) during training.

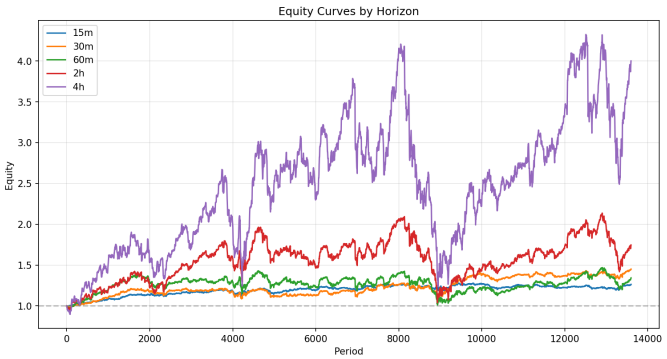


Fig. 3. Equity curves by horizon on test set. Strategy: Long (short) when median prediction positive (negative), position size inversely proportional to interval width. The 4h horizon achieves 300% total return with Sharpe 3.31.

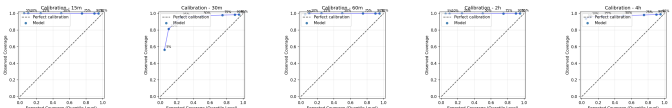


Fig. 4. Calibration reliability diagrams by horizon. Expected coverage (x-axis) versus observed coverage (y-axis). Points below diagonal indicate overconfident predictions; above indicates underconfident. The 30m horizon shows the best calibration.