Modality-Aware Transformers for Financial Time Series Forecasting: an empirical study on the S&P 500 using financial news

Brian Ramesh
brian.ramesh@ensae.fr

Audric Sicard
audric.sicard@ensae.fr

Abstract

Financial return prediction increasingly combines numerical market data with unstructured textual information such as corporate news. While recent deep learning models incorporate multiple modalities, many rely on simple fusion strategies that overlook modality-specific structure and temporal dynamics. We implement and evaluate a Modality-Aware Transformer (MAT) architecture [4] for equity return forecasting, using separate encoders for numerical time series and textual news followed by modality-aware fusion, and compare it against a strong canonical Transformer baseline. Using a U.S. S&P 500 equity dataset from 2010 to 2023, integrating market and sentiment based textual features under a strict point-in-time pipeline, we evaluate models via walk-forward validation. Results show that MAT yields modest improvements in short-horizon cross-sectional ranking performance, particularly at the one-day horizon, while error-based accuracy remains comparable or slightly worse, highlighting the noisy nature of news-driven signals. Overall, text-aware architectures provide limited but targeted benefits when ranking accuracy is prioritized over point forecasting. Code is available at this link.

1 Introduction

Financial markets generate heterogeneous data combining numerical time series with unstructured textual information such as news and corporate disclosures. Prior work (Loughran and McDonald [6]) shows that textual content provides predictive signals for asset prices beyond historical returns alone, notably through sentiment and firm-specific information.

Traditional econometric approaches typically treat numerical and textual data separately, limiting their ability to capture nonlinear dynamics and cross-modal interactions. Recent advances in deep learning, particularly the Transformer architecture (Vaswani et al. [7]) , enable modeling of long-range dependencies and have been applied to financial time series and text. Neural models combining news and market data further highlighted the relevance of textual information for forecasting (Ding et al. [2]; Heaton et al. [5]). However, most multimodal approaches rely on simplistic fusion strategies that ignore modality-specific structure.

The Modality-Aware Transformer (MAT) addresses this limitation by conditioning attention mechanisms on modality identity (Emami et al. [4]). The original study evaluates MAT in settings with densely observed and naturally aligned modalities, such as U.S. interest rate forecasting using macroeconomic time series and official Federal Reserve reports.

In this report, we evaluate MAT in a more challenging and realistic setting: daily stock return prediction using firm-level numerical data and financial news. While numerical features are regularly observed, textual information is sparse, irregular, and noisily aligned. Rather than proposing a new architecture, our goal is to conduct a controlled empirical comparison to investigate whether modality-aware attention offers tangible improvements over standard fusion strategies.

2 Background and method

2.1 Transformer models for sequential data

The Transformer, introduced by Vaswani et al. [7], is a neural network architecture designed to model sequential data using attention mechanisms instead of recurrence. Given an input sequence of length T, each element is mapped to a d-dimensional vector representation and augmented with positional information to encode temporal order. The model then updates each representation by allowing it to attend to other elements in the sequence.

The core operation of the Transformer is self-attention. Given a sequence of embeddings

$$Z = (z_1, \dots, z_T), \quad z_t \in \mathbb{R}^d,$$

each embedding is linearly projected into Query, Key, and Value representations:

$$q_t = W_Q z_t, \qquad k_t = W_K z_t, \qquad v_t = W_V z_t, \tag{1}$$

where $W_Q, W_K, W_V \in \mathbb{R}^{d \times d}$ are learnable parameter matrices.

Self-attention updates the representation at position i as

$$\mathrm{Attn}(z_i) = \sum_{j=1}^{T} \alpha_{ij} v_j, \qquad \alpha_{ij} = \mathrm{softmax}_j \left(\frac{q_i^\top k_j}{\sqrt{d}} \right), \tag{2}$$

where the softmax is taken over the index j. The attention weights α_{ij} quantify the relevance of element j when updating the representation at position i. By adaptively aggregating information across time, self-attention enables the model to capture long-range temporal dependencies, making Transformers well suited for both time series and text modeling.

2.2 Why standard Transformers struggle with multimodal financial data

In many real-world applications, data arise from multiple modalities. In finance, numerical time series describe historical market behavior, while textual data such as news articles convey semantic information about discrete events. A common approach to multimodal modeling is to concatenate representations from all modalities and apply a standard Transformer architecture. This strategy implicitly assumes that all inputs share similar structure and should interact uniformly. In practice, however, numerical and textual data differ substantially in scale, noise, sparsity, and semantic content. Treating them identically may dilute modality-specific information and lead to poorly structured cross-modal interactions. In particular, sparse and irregularly aligned textual signals may dominate or be overwhelmed by dense numerical features, resulting in suboptimal fusion and reduced interpretability. These limitations motivate architectures that explicitly account for modality heterogeneity while still enabling flexible information exchange across modalities.

2.3 Modality-Aware-Transformer (MAT)

Following Emami et al. [4], we adopt a MAT architecture tailored to a two-modality setting consisting of numerical and textual data. Let

$$X^{(n)} = (x_1^{(n)}, \dots, x_T^{(n)}) \quad \text{and} \quad X^{(t)} = (x_1^{(t)}, \dots, x_T^{(t)})$$

denote the numerical and textual input sequences, respectively. We assume that textual representations are temporally aligned to the numerical time grid, e.g. by daily aggregation.

Our implementation follows the general MAT framework while adopting a simplified instantiation adapted to financial time series. The encoder is composed of two parallel modality-specific streams and comprises three main components: feature-level attention, intra-modal temporal self-attention, and inter-modal cross-attention.

Feature-level attention. Before projection to the model dimension, we apply a feature-wise attention mechanism independently within each modality to reweight input variables. For a modality-specific input vector $x_t \in \mathbb{R}^F$ at time t, we compute

$$w_t = \mathrm{softmax}\big(g(x_t)\big), \qquad x'_t = x_t \odot w_t, \tag{3}$$

where $g(\cdot) : \mathbb{R}^F \to \mathbb{R}^F$ is a small multilayer perceptron and $\odot$ denotes elementwise multiplication. The resulting weights emphasize informative features while downweighting less relevant ones and can be interpreted as time-dependent feature importance scores.

Intra-modal temporal self-attention. The reweighted sequences are then projected to a common model dimension and processed independently by Transformer encoder blocks. For a given modality, this corresponds to applying standard self-attention over time:

$$u_t^{(m)} = \sum_{s=1}^{T} \alpha_{ts}^{(m)} \, v_s^{(m)}, \qquad \alpha_{ts}^{(m)} = \mathrm{softmax}_s \left(\frac{q_t^{(m)\top} k_s^{(m)}}{\sqrt{d}} \right), \tag{4}$$

where $m \in \{n, t\}$ indexes the modality, and $(q_t^{(m)}, k_s^{(m)}, v_s^{(m)})$ are query, key, and value representations derived from modality-specific projections. This step captures temporal dependencies within each modality independently.

Inter-modal cross-attention. To enable information exchange across modalities, we introduce bidirectional cross-attention between the two streams. Numerical representations attend to textual representations, and vice versa:

$$\tilde{u}_t^{(n)} = \sum_{s=1}^{T} \beta_{ts}^{(n \leftarrow t)} \, u_s^{(t)}, \qquad \tilde{u}_t^{(t)} = \sum_{s=1}^{T} \beta_{ts}^{(t \leftarrow n)} \, u_s^{(n)}, \tag{5}$$

with attention weights defined analogously to self-attention. Specifically, the coefficient $\beta_{ts}^{(n \leftarrow t)}$ quantifies the relevance of the textual information at step s for the numerical representation at step t. This bidirectional mechanism enables structured cross-modal interactions while preserving modality-specific representations, rather than collapsing all inputs into a single homogeneous sequence. A schematic overview of the architecture is provided in Appendix 6.

2.4 Autoregressive decoding for forecasting

Both the canonical Transformer and MAT adopt an encoder–decoder architecture. The encoder maps input sequences to latent representations, which are then used by an autoregressive decoder to generate forecasts. At time step t, the decoder predicts

$$\hat{y}_t = f_\theta \Big(y_{1:t-1}, \{ \tilde{u}_{1:T}^{(n)}, \tilde{u}_{1:T}^{(t)} \} \Big),$$

where autoregression is enforced via causal masking in the decoder self-attention. During training, we use teacher forcing, feeding the decoder with the true past targets instead of its own predictions to ease optimization , which stabilizes optimization and accelerates convergence. At evaluation time, predictions are generated via autoregressive rollout.

The key architectural difference lies in encoder–decoder attention. The canonical Transformer attends to a single encoder memory, whereas MAT attends separately to modality-specific encoder outputs, which are fused before the feed-forward block. This preserves modality-aware conditioning while maintaining an identical autoregressive decoding scheme across models.

3 Empirical Analysis and Results

In this section, we evaluate the predictive capabilities of the MAT against a strong canonical Transformer baseline. We begin by detailing the construction of our multimodal dataset, which combines high-dimensional numerical market data with unstructured financial news. We then describe the experimental setup, including the training strategy, the optimization protocol, the specific model hyperparameters and the high-performance computing environment used for training. Finally, we present the statistical performance metrics (Information Coefficient (IC), Rank IC, Hit Rate, MAE, MSE B), analyzing the extent to which modality-aware mechanisms improve forecasting accuracy in a noisy financial context. This analysis is done throughout different forecast horizons ($H \in [1, 10]$) with a focus on IC and then assesses the news coverage impact on the IC.

3.1 Dataset and preprocessing

We construct a multimodal dataset spanning January 2010 to December 2023 over the S&P 500 universe. It combines daily market data, fundamental ratios, macroeconomic indicators, and firm-level textual news. All features are processed under a strict point-in-time pipeline to prevent look-ahead bias.

Numerical data. Market data are sourced from CRSP and include daily returns, volatility measures based on the high–low range, liquidity proxies, and cross-sectional market capitalization ranks. Fundamental variables are obtained from WRDS/Compustat and aligned using the `public_date` to ensure availability at decision time, with missing values imputed using sector and global medians. Macroeconomic indicators are collected from FRED, with lower-frequency series shifted forward to account for reporting delays and normalized prior to modeling.

Textual data. Firm-level news is sourced from FNSPID [3] and processed using the pre-trained FinBERT model [1]. For each article, we extract contextual embeddings and sentiment scores, which are aggregated at the stock–date level. We retain summary embedding statistics, sentiment polarity measures, the logarithm of the daily article count, and a binary indicator flagging the presence of news.

Prediction target. The target variable is a volatility-scaled excess forward return,

$$y_{i,t} = \text{clip}\left(\frac{r_{i,t+1} - \mu_{mkt,t+1}}{\sigma_{i,t}}, -5, 5\right),$$

where $r_{i,t+1}$ denotes the close-to-close log return, $\mu_{mkt,t+1}$ the cross-sectional mean return, and $\sigma_{i,t}$ the realized daily volatility. This normalization stabilizes training across heterogeneous assets and frames the task as risk-adjusted return prediction.

Final feature set. The final dataset consists of 21 numerical features, 7 textual features, and one prediction target. A full description of all variables is provided in Appendix C.

3.2 Experimental setup

We compare two primary architectures: the canonical Transformer, which concatenates all features into a single sequence, and the MAT, which preserves modality separation while allowing structured interactions.

Training strategy. Financial time series are non-stationary, meaning market regimes (e.g., bull runs, high-volatility crashes) shift over time. A standard random train-test split would leak future information and fail to account for concept drift. Instead, we employ a strict **walk-forward validation** scheme. The sample period (2010–2023) is partitioned into annual blocks. At each step, the model is trained on a rolling window of five years, validated on the subsequent year, and evaluated on the following year. This "5–1–1" protocol preserves temporal ordering while allowing periodic retraining to adapt to changing market conditions.

Optimization protocol. Model optimization balances numerical stability with economically meaningful predictive performance. Training is performed using the Huber loss ($\delta = 1.0$), which is robust to the heavy-tailed nature of financial returns by interpolating between squared and absolute error penalties. Model selection relies on the information coefficient (IC), which directly evaluates the model's cross-sectional ranking ability. A model can minimize MSE by simply predicting the mean (zero), which is useless for trading. The IC explicitly measures the model's ability to rank assets correctly. Early stopping is applied based on the validation IC: training is terminated if the IC does not improve for five consecutive epochs.

3.3 Experimental Setup

Hyperparameters. To ensure a fair comparison, both models share the same depth, width, and optimization protocol. The specific configuration parameters are detailed in Table 1.

Table 1: Hyperparameter configuration. We utilize a large batch size (8,192) and a OneCycle scheduler to stabilize gradients on the noisy financial dataset.

Parameter	Value
Architecture	
Input Dimension (d_{model})	128
Feed-Forward Dimension (d_{ff})	512
Number of Heads (H)	4
Number of Layers (L)	2
Dropout Rate	0.2
Optimization	
Optimizer	AdamW
Scheduler	OneCycle (10% Warmup)
Peak Learning Rate	7×10^{-4}
Batch Size	8,192
Lookback Window (T)	60 days
Forecast Horizon	1–10 days

Training environment. Experiments were conducted on a high-performance computing cluster equipped with an NVIDIA A100 Tensor Core GPU (80 GB VRAM). The large memory capacity enabled the use of a batch size of 8,192, which contributes to stabilizing Transformer training under noisy financial data. To maximize data throughput, we employed persistent data loader workers with a prefetch factor of four, allowing efficient overlap between data loading and GPU computation.

3.4 Statistical Results

3.4.1 Pooled Horizon Analysis

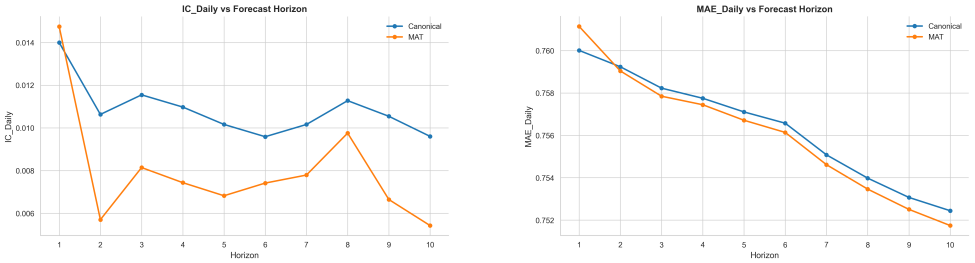
We first evaluate aggregate performance by pooling daily predictions across all forecast horizons ($H \in [1, 10]$). Table 2 reports the assessment of predictive accuracy and cross-sectional ranking ability. Across pooled horizons, the canonical Transformer slightly outperforms the MAT model on ranking-based metrics, achieving higher mean IC and rank IC. Directional metrics such as hit rate and top-quintile precision exhibit a similar, though economically modest, advantage for the canonical model. Conversely, error-based metrics (MAE and MSE) are nearly identical, indicating comparable point forecast accuracy. For both models, the daily IC displays high volatility relative to its mean, reflecting the noisy nature of short-horizon financial prediction.

Table 2: **Pooled Statistical Performance (Daily Average).** Metrics are averaged over 1,915 testing days. Standard deviations represent daily volatility. P-values for the DM test are against a zero-return benchmark, see Appendix D.

Metric	canonical	MAT
Information Coefficient (IC)	0.0094 ± 0.0602	0.0085 ± 0.0586
IC t-statistic	6.82	6.33
Rank IC (Spearman)	0.0091 ± 0.0609	0.0075 ± 0.0559
Rank IC t-statistic	6.57	5.85
Hit Rate	50.49%	50.28%
Precision (Top 20%)	50.97%	50.89%
Mean Absolute Error (MAE)	0.7563 ± 0.1067	0.7561 ± 0.1057
Mean Squared Error (MSE)	1.1472 ± 0.3193	1.1448 ± 0.3181
DM Test p-value	< 0.001	< 0.001

3.4.2 Horizon analysis and signal decay

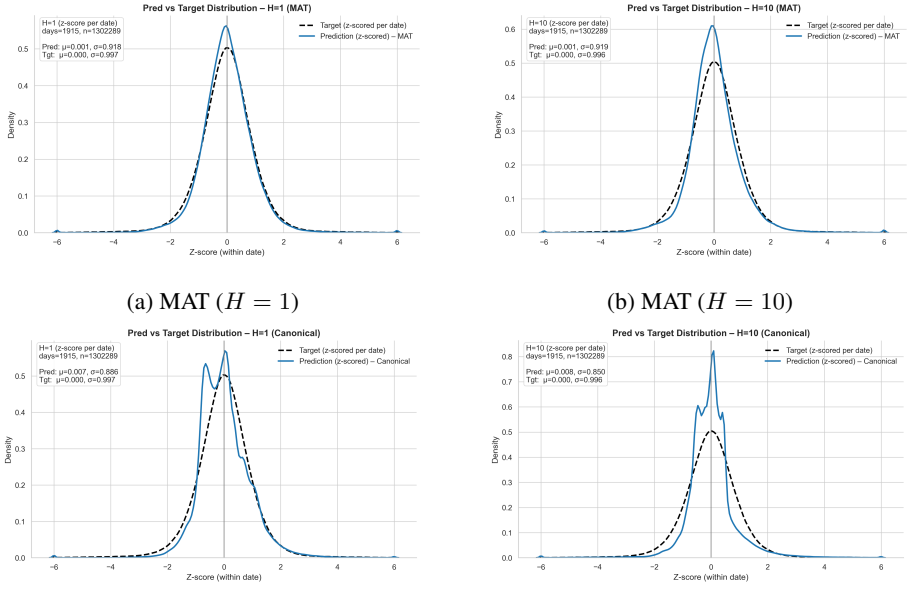
Decomposing performance by forecast horizon reveals marked architectural differences. We analyze horizon effects using the IC for ranking and MAE for point accuracy. About ranking performance and IC decay, Figure 1(a) shows a sharp contrast in signal persistence. MAT achieves its highest IC at $H = 1$ but experiences a rapid collapse at $H = 2$, suggesting news-driven signals are highly transient. The canonical Transformer exhibits a more gradual decay, suggesting a reliance on persistent numerical factors. Figure 1(b) reports MAE across horizons. Despite ranking decay, MAE decreases as the horizon increases for both models. This "Decreasing Error Paradox" highlights the tension between ranking and error minimization: as signals weaken, models shrink predictions toward zero to limit large absolute errors.



(a) IC decay by horizon. (b) MAE decay by horizon.

Figure 1: **Forecast horizon effects.** IC and MAE as a function of forecast horizon.

A focus on predictions distributions gives more insight into those metrics decay. Figure 2 illustrates the underlying mechanism. At longer horizons, the canonical model produces concentrated distributions centered at zero (explicit shrinkage). MAT maintains a stable spread, leading to active but noisier predictions at short horizons, but conservative forecasts that limit large errors at longer horizons.



(a) MAT ($H = 1$) (b) MAT ($H = 10$)

(c) canonical ($H = 1$) (d) canonical ($H = 10$)

Figure 2: **Distributional dynamics.** Prediction–target distributions. As the horizon increases, canonical produces more concentrated predictions (shrinkage), while MAT maintains stable dispersion.

3.4.3 Horizon one analysis: the precision-stability trade-off

We focus on horizon 1 ($H = 1$), which captures the market's immediate reaction to new information and constitutes the primary operating regime of the MAT. Table 3 reports the corresponding performance metrics.

Table 3: **Horizon 1 performance comparison.** Information Coefficient (IC), Mean Absolute Error (MAE), and statistical tests comparing the canonical Transformer and MAT at the one-day forecast horizon.

Metric	canonical	MAT	Difference
IC (Alpha)	0.0140	**0.0147**	+5.3%
MAE (Error)	**0.7600**	0.7611	+0.15%
DM Test p-value (Errors)	0.035		*Significant*
IC Diff p-value (Alphas)	0.868		*Not Significant*

At the one-day horizon, the two architectures exhibit a clear trade-off between ranking performance and error stability. MAT achieves a higher IC, indicating stronger short-term alpha capture, while the canonical Transformer attains a lower MAE, reflecting more conservative forecasts. The Diebold–Mariano test confirms that forecast error dynamics differ significantly ($p = 0.035$), whereas the IC difference is not statistically significant ($p = 0.868$, see see Appendix E).

The canonical model favors error minimization through reliance on persistent numerical signals, while MAT exploits transient textual information to generate more aggressive short-term predictions. This strategy increases variance and slightly worsens MAE but can improve contemporaneous ranking performance. Figure 3 shows a modest upward shift in MAT's IC distribution at $H = 1$, although substantial overlap across models prevents statistical significance, underscoring the noisy nature of short-horizon financial signals.

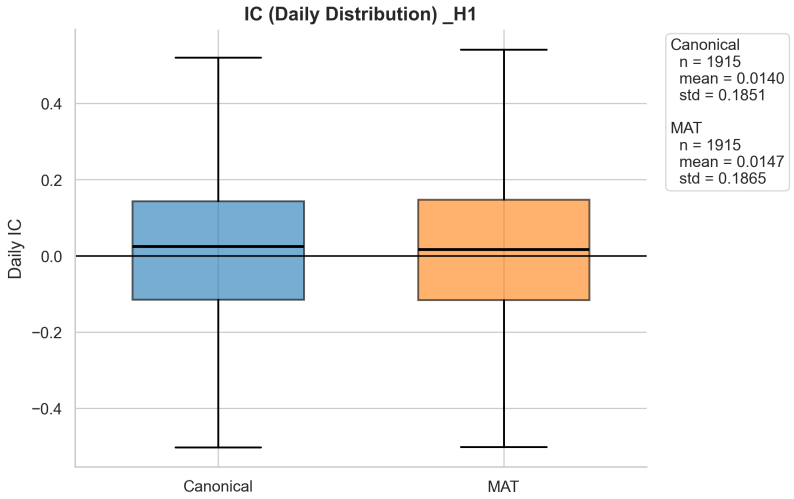


Figure 3: **Daily IC distribution at $H = 1$.** Boxplots of daily Information Coefficients for the canonical Transformer and MAT models. The MAT distribution (orange) exhibits a slightly higher median and positive bias compared to the canonical baseline (blue), though the variance remains high for both.

3.4.4 Modality impact and information density analysis

The effectiveness of multimodal architectures such as MAT depends critically on the availability of the secondary modality. We therefore examine the temporal evolution of news coverage in our dataset to contextualize the cross-sectional results.

News intensity varies substantially over time, with an average daily coverage of 31.9% (and a standard-deviation of 11.4 percentage points), ranging from 5.8% to a peak of 71.9%. This heterogeneity provides a suitable setting to assess whether modality-aware architectures deliver systematic gains when textual information is available.

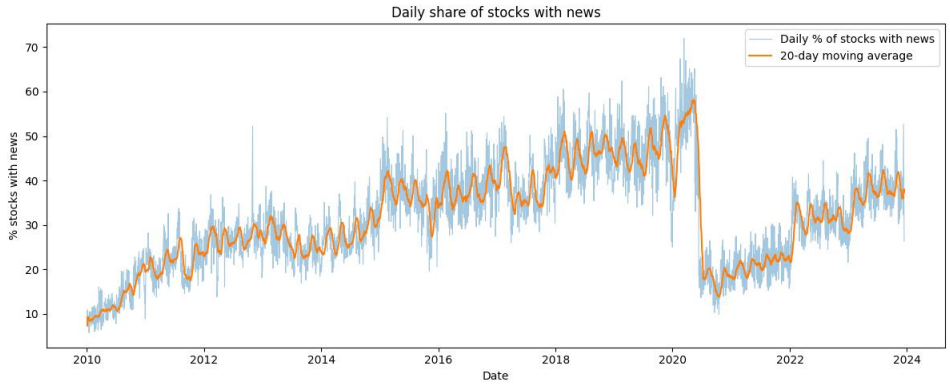


Figure 4: **Temporal evolution of news coverage.** Daily share of S&P 500 stocks with at least one news article (blue) and its 20-day moving average (orange).

The News efficiency gap. To assess whether the predictive advantage of modality-aware architectures is structural, we decompose performance by the presence of firm-specific news over the 10-day forecast window. Figure 5 reveals a clear bifurcation of the S&P 500 universe into two predictability regimes. Across nearly all horizons, both models exhibit a pronounced performance gap between stocks with news coverage and those without. At the one-day horizon, peak performance is achieved in the absence of news, with MAT and the canonical model reaching ICs of 0.0195 and 0.0190, respectively. In contrast, performance on news days collapses by more than 65% for both architectures. This pattern indicates that firm-specific news rapidly increases market efficiency, leaving little exploitable signal for daily-frequency models. Consequently, the modest short-horizon advantage of MAT does not reflect a structural edge in news-rich environments.

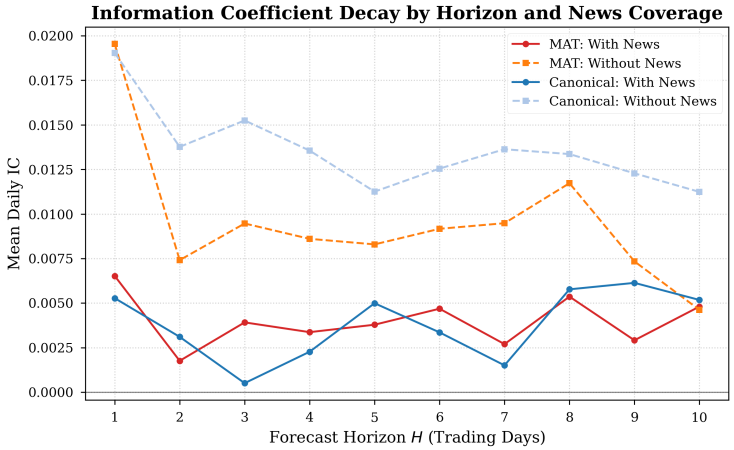


Figure 5: **IC decay by forecast horizon and news presence.** Mean Information Coefficient as a function of the forecast horizon, reported separately for assets with (solid lines) and without (dashed lines) firm-specific news.

Architectural specialization. Table 4 highlights a clear horizon- and regime-dependent difference between the two architectures. When firm-specific news is present, MAT outperforms the canonical Transformer at the one-day horizon ($H = 1$), indicating a more effective exploitation of short-lived textual signals. However, this advantage disappears rapidly beyond the immediate horizon.

In contrast, in the absence of news, the canonical model exhibits greater robustness. Its ranking performance remains stable across longer horizons, while MAT's IC decays more sharply. This suggests that the canonical architecture is better suited for capturing persistent numerical patterns that dominate low-information environments.

Table 4: **Mean daily IC by forecast horizon and news presence.** Results are reported separately for stocks with and without firm-specific news coverage.

	With News		Without News	
Horizon	**MAT IC**	**Can IC**	**MAT IC**	**Can IC**
$H = 1$	**0.0065**	0.0053	0.0195	0.0190
$H = 2$	0.0018	0.0031	0.0074	**0.0138**
$H = 5$	0.0038	0.0050	0.0083	**0.0113**
$H = 10$	0.0048	0.0052	0.0046	**0.0112**

These findings suggest that the MAT model acts as a "Sniper," specialized for capturing the immediate, transient alpha embedded in textual shocks. However, for capturing the slower-moving, persistent numerical factors that dominate quiet market periods, the canonical Transformer's simpler integration of features provides a more reliable long-term signal.

4 Conclusion

This report evaluated whether explicitly preserving modality-specific structure improves multimodal equity return forecasting. We implemented a Modality-Aware Transformer (MAT) [4] and compared it to a strong canonical Transformer baseline that fuses numerical and textual inputs via simple feature concatenation. Using a strict point-in-time dataset on the S&P 500 from 2010 to 2023 and a walk-forward validation protocol, we assessed performance across forecast horizons using both error-based metrics (MAE) and ranking metrics (IC/Rank IC).

Our results indicate that modality-aware attention yields *targeted* benefits rather than broad improvements. MAT exhibits its strongest advantage at the one-day horizon, where it can better exploit transient news-driven signals and slightly improves cross-sectional ranking performance. However, this advantage decays quickly beyond the immediate horizon, and pooled results favor the canonical model on ranking metrics. Error-based accuracy remains comparable across architectures, highlighting that small gains in ranking can coexist with similar (or slightly worse) point forecast error in noisy financial environments.

A key empirical finding is that predictability is strongly regime-dependent: performance is substantially higher on stock-days without firm-specific news, while both models struggle when news is present, consistent with rapid information incorporation. Overall, our findings suggest that MAT-style architectures can be useful when the objective prioritizes short-horizon ranking, but that their impact is constrained by the sparsity and efficiency of public news signals at daily frequency.

Finally, because statistical metrics do not directly translate into economic value, we provide an auxiliary portfolio-based diagnostic in Appendix F, which converts model forecasts into long–short signals to assess stability, turnover, and cross-sectional separation. This analysis supports the same qualitative conclusion: MAT can generate a stronger short-term ranking signal, but its economic relevance remains sensitive to trading frictions and regime shifts.

Future work should focus on improving the temporal alignment of textual information (e.g., intraday timestamps), enriching the news corpus, and benchmarking against stronger baselines and ablations (numerical-only, text-only, and alternative fusion mechanisms) to better characterize when multimodal attention yields robust and economically meaningful gains in this framework.

References

[1] Dogu Araci. Finbert: Financial sentiment analysis with pre-trained language models, 2019. URL https://arxiv.org/abs/1908.10063.

[2] Xiao Ding, Yue Zhang, Ting Liu, and Junwen Duan. Deep learning for event-driven stock prediction. In *Ijcai*, volume 15, pages 2327–2333, 2015.

[3] Zihan Dong, Xinyu Fan, and Zhiyuan Peng. Fnspid: A comprehensive financial news dataset in time series, 2024.

[4] Hajar Emami, Xuan-Hong Dang, Yousaf Shah, and Petros Zerfos. Modality-aware transformer for financial time series forecasting, 2024. URL https://arxiv.org/abs/2310.01232.

[5] James B Heaton, Nick G Polson, and Jan Hendrik Witte. Deep learning for finance: deep portfolios. *Applied Stochastic Models in Business and Industry*, 33(1):3–12, 2017.

[6] Tim Loughran and Bill McDonald. When is a liability not a liability? textual analysis, dictionaries, and 10-ks. *The Journal of finance*, 66(1):35–65, 2011.

[7] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

Appendix

A Modality-Aware Transformer (MAT) Architecture

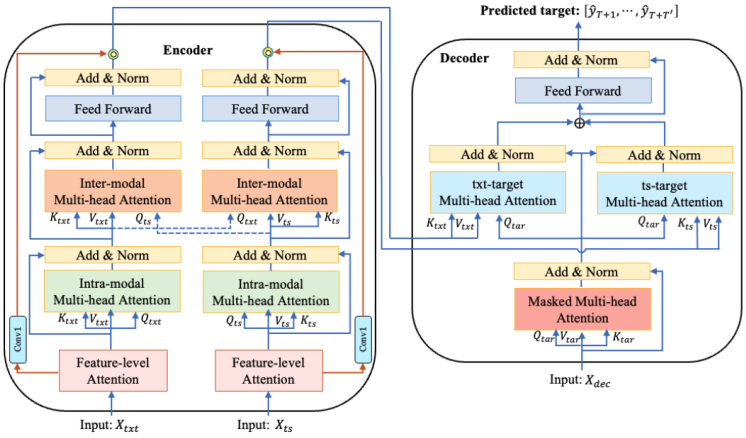


Figure 6: The Modality-aware Transformer (MAT) architecture (reproduced from Emami et al. [4]).

B Evaluation Metrics

This appendix defines the evaluation metrics used to assess forecast accuracy and predictive relevance.

B.1 Mean Absolute Error (MAE)

The Mean Absolute Error (MAE) measures the average magnitude of forecast errors, without regard to their direction. Given realized values y_t and predictions $\hat{y}_t$ for $t = 1, \ldots, T$, it is defined as:

$$\text{MAE} = \frac{1}{T} \sum_{t=1}^{T} |y_t - \hat{y}_t| \,.$$

MAE is easy to interpret and relatively robust to outliers, as it penalizes all deviations linearly. It captures the typical absolute deviation between forecasts and realizations.

B.2 Mean Squared Error (MSE)

The Mean Squared Error (MSE) provides a measure of the average squared difference between predictions and realizations. It is defined as:

$$\text{MSE} = \frac{1}{T} \sum_{t=1}^{T} (y_t - \hat{y}_t)^2 \,.$$

Unlike MAE, MSE penalizes larger errors more heavily due to the squaring of the residual terms. In financial modeling, it is particularly sensitive to outliers and extreme deviations, making it a useful metric for assessing the stability of point forecasts.

B.3 Information Coefficient (IC)

The Information Coefficient (IC) evaluates the cross-sectional association between model forecasts and realized outcomes. At each time t, it is computed as:

$$\text{IC}_t = \text{corr}(\hat{y}_{t,i},\, y_{t,i}) \,,$$

where i indexes assets in the cross-section. The reported IC is typically the time average of IC_t.

Unlike MAE and MSE, the IC assesses directional and ranking accuracy rather than error magnitude. A positive IC indicates that higher predicted values are associated with higher realized outcomes, reflecting the model's ability to correctly rank assets.

B.4 Rank Information Coefficient (Rank IC)

The Rank IC is a non-parametric measure of the association between predictions and realizations. It is defined as the Spearman rank correlation between the predicted values and the actual outcomes at time t:

$$\text{Rank IC}_t = \rho(\text{rank}(\hat{y}_{t,i}), \text{rank}(y_{t,i}))\,.$$

By utilizing the ranks of the variables rather than their raw values, the Rank IC is robust to non-linearities and the influence of outliers. This metric is essential for evaluating the effectiveness of cross-sectional ranking strategies, where the order of assets is more important than the specific predicted magnitude.

B.5 Hit rate

The Hit Rate (or Directional Accuracy) measures the proportion of forecasts that correctly predict the sign of the realized outcome. For a sample of T observations across assets, it is defined as:

$$\text{Hit Rate} = \frac{1}{N}\sum_{j=1}^{N}\mathbb{I}(\text{sgn}(\hat{y}_j) = \text{sgn}(y_j))\,,$$

where $\mathbb{I}$ is the indicator function. The Hit Rate assesses the binary classification performance of the model, specifically its ability to distinguish between positive and negative returns. A Hit Rate significantly above 50% indicates that the model possesses directional predictive skill beyond random chance.

C Input Features

Table 5: Complete list of input features and prediction target.

Category	Feature Name	Description
Market	`mkt_log_ret`	Daily log return (t vs $t-1$)
	`mkt_volatility`	Log Parkinson volatility (estimator using High/Low)
	`mkt_mom_1m`	1-month cumulative momentum
	`mkt_mom_3m`	3-month cumulative momentum
	`mkt_liq_risk`	Log Liquidity Risk (Bid-Ask Spread proxy)
	`mkt_turnover`	Log Daily Share Turnover (Volume / Shares Out)
	`mkt_rel_vol`	Log Relative Volume (vs 20-day MA)
	`mkt_drawdown`	Log Drawdown from 1-year High
	`mkt_cap_rank`	Cross-sectional Rank of Market Capitalization
Ratio	`ratio_pb`	Log Price-to-Book Ratio
	`ratio_ey`	Earnings Yield (Inverse P/E)
	`ratio_roe`	Return on Equity (Capped)
	`ratio_de`	Log Debt-to-Equity Ratio
	`ratio_div_yield`	Dividend Yield
Macro	`macro_risk_free`	10-Year Treasury Rate
	`macro_yield_curve`	Yield Curve Slope (10Y - 2Y Spread)
	`macro_vix`	Log VIX Index (Market Volatility)
	`macro_unemp_rate`	US Unemployment Rate
	`macro_unemp_delta`	Monthly Change in Unemployment
	`macro_cpi_yoy`	CPI Inflation (Year-over-Year)
	`macro_ppi_yoy`	PPI Inflation (Year-over-Year)
Text	`emb_mean`	768-d average FinBERT embedding
	`sent_score_mean`	Avg. polarity score ($P_{pos} - P_{neg}$)
	`sent_pos_mean`	Average probability of positive sentiment
	`sent_neg_mean`	Average probability of negative sentiment
	`sent_score_std`	Standard deviation of sentiment (opinion divergence)
	`log_n_news`	Logarithm of daily article count
	`has_news`	Binary flag (1 if news exists, 0 otherwise)
Target	`target`	Volatility-Scaled Excess Return (Sharpe ratio Proxy)

D Diebold–Mariano test

The Diebold–Mariano (DM) test is a statistical procedure used to compare the predictive accuracy of two forecasting models. It tests the null hypothesis that both models have equal expected predictive loss under a given loss function (e.g., MAE or MSE). The test is widely used in time-series forecasting as it explicitly accounts for serial correlation in the sequence of forecast losses.

Setup. Let y_t denote the observed value at time t, and let $\hat{y}_t^{(1)}$ and $\hat{y}_t^{(2)}$ be the corresponding forecasts produced by model 1 and model 2, respectively. Define the forecast errors

$$e_t^{(i)} = y_t - \hat{y}_t^{(i)}, \qquad i \in \{1, 2\}.$$

Given a loss function $L(\cdot)$ (e.g., squared or absolute error), the loss differential is defined as

$$d_t = L\left(e_t^{(1)}\right) - L\left(e_t^{(2)}\right).$$

A positive value of $\mathbb{E}[d_t]$ indicates that model 2 has lower expected loss and thus superior predictive accuracy, and vice versa.

Test statistic. Let $\bar{d} = T^{-1} \sum_{t=1}^{T} d_t$ denote the sample mean of the loss differentials. The DM test statistic is given by

$$DM = \frac{\bar{d}}{\sqrt{\widehat{\mathrm{Var}}(\bar{d})}},$$

where $\widehat{\mathrm{Var}}(\bar{d})$ is a heteroskedasticity- and autocorrelation-consistent estimate of the variance of $\bar{d}$. In practice, this variance is commonly estimated using a Newey–West estimator. For h-step-ahead forecasts, the estimator accounts for serial correlation in $\{d_t\}$ up to lag $h - 1$.

Under the null hypothesis

$$H_0 : \mathbb{E}[d_t] = 0,$$

the DM statistic is asymptotically distributed as a standard normal random variable.

Interpretation. A small p-value leads to rejection of the null hypothesis of equal predictive accuracy, indicating that one forecasting model significantly outperforms the other under the chosen loss function. In our empirical analysis, we compute DM p-values relative to a naive zero-forecast baseline in order to assess whether the proposed models provide statistically significant predictive gains.

The DM test depends on the choice of loss function and may be sensitive to outliers, particularly when squared error losses are used. Moreover, the test relies on asymptotic theory, and its finite-sample performance can be poor when the evaluation sample is short or when loss differentials exhibit strong dependence. Results should therefore be interpreted with caution in small samples.

E Paired t-test for IC difference

The paired t-test is a statistical procedure used to determine whether the mean difference between two sets of paired observations is zero. In our framework, we use this test to compare the ranking performance of two models by evaluating the statistical significance of the difference in their daily Information Coefficients (IC).

Setup. Let $\mathrm{IC}_t^{(1)}$ and $\mathrm{IC}_t^{(2)}$ denote the cross-sectional Information Coefficients for model 1 and model 2, respectively, calculated at time t. The Information Coefficient at each time step is defined as the Pearson correlation between the model's predictions and the realized targets for that cross-section.

The IC differential for each period is defined as:

$$\delta_t = \mathrm{IC}_t^{(1)} - \mathrm{IC}_t^{(2)}.$$

The null hypothesis H_0 posits that there is no systematic difference in the ranking ability of the two models:

$$H_0 : \mathbb{E}[\delta_t] = 0.$$

A positive mean value of δ_t suggests that model 1 provides superior cross-sectional ranking compared to model 2, while a negative value suggests the opposite.

Test statistic. Let $\bar{\delta} = T^{-1} \sum_{t=1}^{T} \delta_t$ be the sample mean of the IC differentials and s_δ be the sample standard deviation of the differentials. The paired t-statistic is computed as:

$$t_{\mathrm{IC}} = \frac{\bar{\delta}}{s_\delta / \sqrt{T}},$$

where T represents the number of days in the evaluation period. Under the assumption that the differentials are independent and identically distributed (i.i.d.) and follow a normal distribution, the statistic follows a t-distribution with $T - 1$ degrees of freedom.

Interpretation and limitations. A p-value lower than a chosen significance level (e.g., 0.05) indicates that the difference in ranking performance between the two models is statistically significant and unlikely to have arisen from random noise. In our results, this test is specifically used to assess whether the MAT architecture's alpha generation is significantly different from the Canonical baseline.

It is important to acknowledge that the paired t-test is not the most robust method for this comparison, as its validity relies on the assumption that the daily observations are independent. In financial time series, daily metrics often exhibit serial correlation or heteroskedasticity. While the paired nature of the test effectively controls for day-specific market shocks that affect both models simultaneously, it does not explicitly account for the temporal dependence of the IC differentials. Consequently, the resulting t-statistics and p-values should be interpreted as indicative of relative performance rather than absolute evidence of long-term superiority.

F Signal Construction and Portfolio Analysis

Statistical metrics such as MAE or Information Coefficient (IC) evaluate predictive accuracy at the observation level, but do not fully capture how well a model separates outperforming from underperforming assets. In practice, economically relevant signals require consistent cross-sectional discrimination rather than small average errors.

In this section, we therefore transform model forecasts into simple long–short portfolio signals. This analysis is intended as a diagnostic tool to assess ranking quality, stability, and signal decay, rather than as a fully realistic trading strategy. All results are reported in gross terms, without transaction costs, to isolate the intrinsic predictive structure of the models.

F.1 Signal construction strategies

The models produce daily return forecasts $\hat{r}_{t+h}$ for horizons $h \in \{1, \ldots, 10\}$, where $\hat{r}_{t+h}$ denotes the predicted close-to-close return realized on day $t + h$. Importantly, these forecasts correspond to individual daily returns and not to cumulative returns over the interval $[t, t + h]$. For instance, if t is a Monday, the horizon $h = 5$ forecast corresponds to the return realized on the following Monday. To evaluate how predictive information evolves across horizons, we transform these forecasts into several portfolio signals. Each signal emphasizes a different trade-off between short-term precision and temporal stability.

1. **H1 only.**
$$S_{i,t}^{H1} = \hat{r}_{i,t+1}$$
This signal isolates immediate next-day predictive accuracy and serves as a benchmark for short-horizon performance.

2. **H1–H5 mean.**
$$S_{i,t}^{H1\text{–}H5} = \frac{1}{5} \sum_{h=1}^{5} \hat{r}_{i,t+h}$$
Averaging predictions over one week smooths idiosyncratic daily noise and captures short-term consistency in the forecasts.

3. **H1–H10 mean.**
$$S_{i,t}^{H1\text{–}H10} = \frac{1}{10} \sum_{h=1}^{10} \hat{r}_{i,t+h}$$
This signal probes longer-horizon stability and assesses whether predictive power persists beyond the immediate time frame.

4. **Smart decay.**
$$S_{i,t}^{SD} = \sum_{h=1}^{5} \frac{1}{h} \hat{r}_{i,t+h}$$
Closer horizons receive higher weight, reflecting increasing uncertainty as the forecast horizon grows while retaining medium-term information.

5. **Conviction filter.**
$$S_{i,t}^{Conv} = \begin{cases} \hat{r}_{i,t+1}, & \text{if } \operatorname{sign}(\hat{r}_{i,t+1}) = \operatorname{sign}(\hat{r}_{i,t+5}) \\ 0, & \text{otherwise.} \end{cases}$$
This non-linear filter retains positions only when short- and medium-horizon forecasts agree in direction, reducing exposure to transient and contradictory signals.

F.2 Portfolio construction and backtest logic

To evaluate the cross-sectional ranking quality of the models, we implement a simple long–short portfolio backtest based on the constructed signals. At each date, assets are ranked according to their predicted returns, and positions are taken in the highest (and lowest) ranked stocks.

The investment universe is restricted to the active constituents of the **S&P 500** at each time step t, ensuring liquidity and consistency over time. To isolate relative stock selection from market-wide movements, all portfolios are constructed to be dollar-neutral, with equal exposure on the long and short sides.

Ranking and allocation. At each time step t, we rank the N_t stocks available in the filtered S&P 500 universe based on the signal $S_{i,t}$. We go long the top decile (Top 10%) and short the bottom decile (Bottom 10%). The weights $w_{i,t}$ are constructed as follows:

- **Long leg:** equal weight among the top 10% of stocks.
- **Short leg:** equal weight among the bottom 10% of stocks.
- **Net exposure:** 0% (dollar neutral).
- **Gross exposure:** 100% (leverage factor of 1.0).

Turnover and transaction costs. As our primary objective is to evaluate the theoretical predictive power and ranking quality of the models rather than to validate a live execution strategy, we assume zero transaction costs. We do not apply slippage or commission penalties to the returns. However, we calculate daily portfolio turnover as a proxy for signal stability:

$$\text{Turnover}_t = \frac{1}{2} \sum_i |w_{i,t} - w_{i,t-1}^{\text{drift}}|$$

where w^{drift} represents the portfolio weights after price evolution but before rebalancing. Consequently, the reported portfolio returns represent the pure gross performance of the selected assets.

F.3 Horizon analysis and signal decay

We assess how stable the predictive signals are as the forecast horizon increases. The goal is to distinguish models that capture persistent patterns from those that mainly react to short-term noise. To this end, we compare portfolio performance across the different horizon-based signals introduced above.

Performance is summarized using the Sharpe Ratio, defined as the average portfolio return divided by its standard deviation. Table 6 reports the results for all models and signal constructions. Three main observations emerge.

- **Short-horizon strength of MAT.** MAT achieves its highest performance when using the one-day signal ($H = 1$), with a Sharpe Ratio of **0.52**. This indicates that the model is particularly effective at extracting high-precision, short-lived signals from multimodal inputs.
- **Effect of horizon averaging.** Signals based on a single-day forecast exhibit high turnover, reflecting frequent portfolio rebalancing and lower signal stability. Averaging predictions over longer horizons substantially reduces turnover. Averaging predictions over longer horizons substantially reduces turnover. For MAT, however, this smoothing also weakens performance, with the Sharpe Ratio declining to 0.17 at $H = 10$. This suggests that MAT's predictive advantage is concentrated at very short horizons.
- **Smoothing requirement of the canonical model.** In contrast, the canonical model performs poorly when trading directly on one-day forecasts (Sharpe 0.18). Its performance improves only when predictions are averaged across multiple horizons, reaching a peak Sharpe of 0.24 for the $H1$–$H5$ signal. This behavior indicates a reliance on more persistent, slowly evolving signals.

Regime shift and 2020 drawdown. Figure 7 visually confirms these dynamics but also highlights a critical failure mode common to both architectures. A sharp, synchronized drawdown is visible across all signals in the first half of 2020. This corresponds to the onset of the COVID-19 pandemic, a regime shift, meaning a sudden change in market behavior that invalidates patterns learned during calmer periods.

Both models fail to predict this crash correctly in the sense that their ranking signals generate large negative portfolio returns. This is likely attributable to the distribution of the training data (2014–2018)

Table 6: Performance metrics for all signal strategies. MAT achieves peak performance with high-precision short-term signals (H1), while canonical benefits from smoothing. Turnover decreases significantly as the horizon extends, reflecting the trade-off between signal agility and execution cost.

Model	Strategy	Ann. Return	Ann. Vol	Sharpe	Max DD	Turnover
canonical	H1 Only	1.70%	9.50%	0.18	-28.94%	123.35
canonical	H1-H5 Mean	2.31%	9.75%	0.24	-25.79%	86.77
canonical	H1-H10 Mean	1.96%	9.74%	0.20	-24.35%	74.86
canonical	Smart Decay	2.73%	9.75%	0.28	-26.92%	96.83
MAT	H1 Only	**5.03%**	9.71%	**0.52**	-21.80%	101.54
MAT	H1-H5 Mean	2.56%	9.89%	0.26	-18.69%	55.53
MAT	H1-H10 Mean	1.69%	9.96%	0.17	-19.94%	42.10
MAT	Smart Decay	3.00%	9.90%	0.30	-18.54%	68.43

and validation data (2019), which covered a prolonged period of relative market stability. Having not encountered such an extreme exogenous shock during training, the models continued to forecast based on stable-regime dynamics, resulting in significant losses during the crash. However, the MAT H1 signal recovers more robustly in the post-2020 period compared to the canonical baselines.

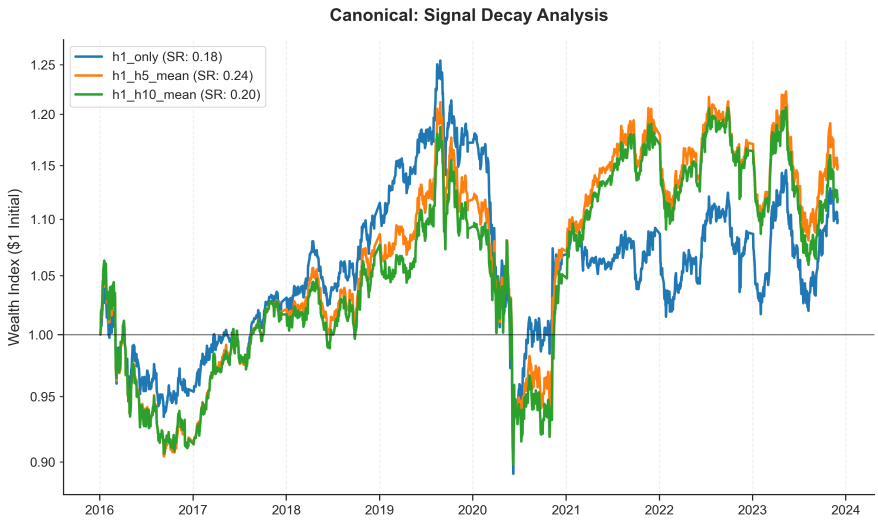
Discriminative power (quintile analysis). Quintile portfolios provide a simple way to visualize ranking quality by grouping assets from best to worst predictions and comparing their realized performance over time. At each date, assets are ranked according to the conviction signal and partitioned into five equally sized cross-sectional portfolios (quintiles), from the highest-ranked (Q1) to the lowest-ranked (Q5). To verify that the model's alpha, understood here as persistent excess performance relative to the cross-sectional average, is not driven by outliers or idiosyncratic bets, we examine the monotonic separation of returns across quintiles. Figure 8 compares the cumulative wealth of portfolios, obtained by compounding daily portfolio returns over time, sorted by the conviction signal for both the MAT and canonical champions.

The **MAT model** (Panel b) reveals a robust, albeit imperfect, stratification. A clear separation exists between the top and bottom tiers: the top quintiles (Q1 and Q2) decouple from the rest of the pack to the upside, while the bottom quintile (Q5, red line) persistently drags performance down, ending significantly below all others. While the ordering is not strictly monotonic, notably, **Q2 finishes slightly above Q1** and **Q4 ends above Q3**, the model successfully identifies a cluster of outperformers versus underperformers, confirming its strong discriminatory power at the extremes of the distribution.

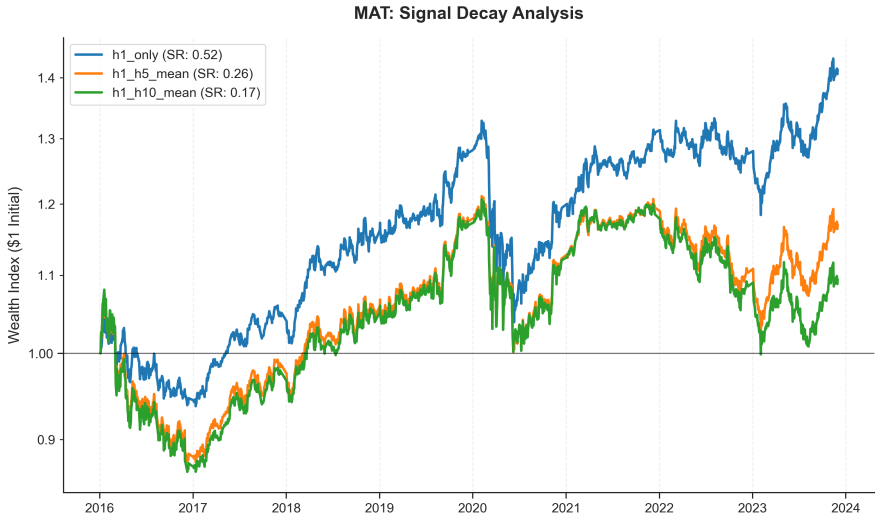
In contrast, the **canonical model** (Panel a) exhibits significantly compressed separation. While Q1 still outperforms Q5, the middle quintiles (Q2, Q3, Q4) are tightly clustered and frequently cross over each other. This lack of distinct ordering suggests that the canonical Transformer struggles to differentiate between "average" and "good" stocks, relying mostly on a crude separation of the extreme tails.

Both models, however, share a common vulnerability: the sharp, synchronized drawdown in early 2020. This confirms that while MAT is superior at relative ranking, neither architecture could circumvent the systemic risk posed by the COVID-19 regime shift.

Long/Short attribution. A common failure mode in machine learning portfolios is "Short-Side Failure," where the model identifies winners but fails to distinguish losers from the market average. Table 7 decomposes the returns by leg. Both models successfully identify underperforming stocks, with Short Leg returns contributing significantly to alpha (recall that a negative raw return on the Short Leg implies a profit for the strategy). However, the MAT Champion distinguishes itself on the **Long Leg**, generating an annualized return of **4.00%** compared to the canonical model's **3.01%**. This indicates that the modality-aware fusion is particularly effective at capturing high-conviction positive sentiment, leading to superior stock selection on the upside.



(a) canonical Transformer

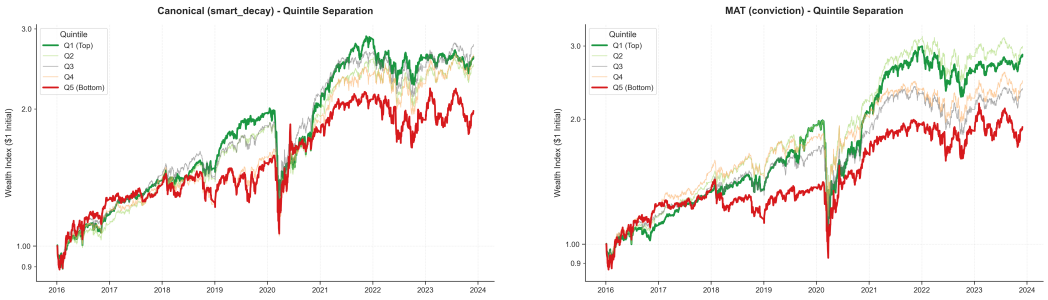


(b) Modality-Aware Transformer (MAT)

Figure 7: **Signal decay analysis.** The canonical model (Top) produces flat curves that require smoothing to generate stable returns. The MAT model (Bottom) exhibits strong alpha generation in the short term ($H = 1$), though the signal strength dilutes with averaging. *Note:* Both models suffer a significant drawdown in mid-2020, likely due to the inability to generalize from the stable training period (2014–2019) to the extreme volatility of the COVID-19 pandemic.

Robustness check. Finally, we evaluate the statistical stability of the strategies using a Stationary Block Bootstrap ($N = 2000$, block size=22 days) to preserve serial correlation. Figure 9 compares the distributions of Sharpe Ratios.

The **MAT Champion** (Panel b) demonstrates superior consistency with a daily **win rate of 52.2%**, compared to 51.8% for the canonical baseline. While the canonical model exhibits higher positive skewness (0.54 vs 0.05), suggesting reliance on occasional extreme outliers, the MAT model displays a more symmetric return profile, indicative of systematic rather than episodic performance.



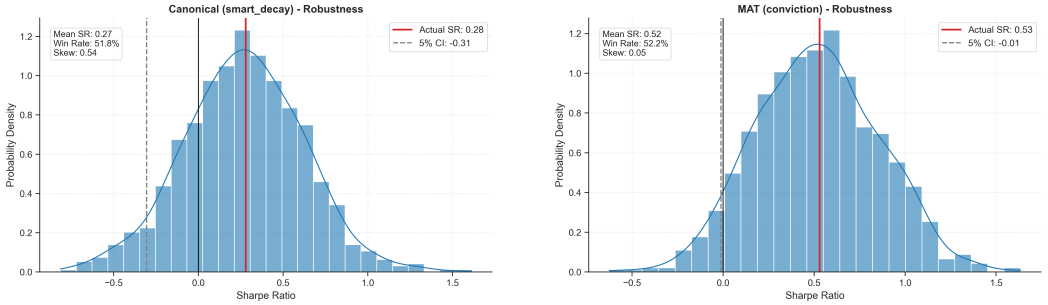
(a) canonical (Smart Decay) (b) MAT (Conviction)

Figure 8: **Discriminative power analysis.** (a) The canonical model shows compressed separation with overlapping middle quintiles, indicating weak ranking ability. (b) The MAT model displays better, monotonic "fanning out" of returns ($Q1 > Q5$), confirming it can sort assets by quality. Note the shared drawdown in 2020 across all quintiles.

Table 7: **Return attribution by Leg.** MAT outperforms on the Long Leg (4.00% vs 3.01%) while maintaining comparable Short Leg performance.

Model	Leg	Ann. Return	Ann. Volatility
canonical *(Smart Decay)*	Long (Top 10%)	3.01%	6.89%
	Short (Bot 10%)	-1.85%	7.62%
MAT *(Conviction)*	Long (Top 10%)	**4.00%**	7.36%
	Short (Bot 10%)	-1.84%	7.18%

Regarding statistical significance, the MAT model's mean bootstrapped Sharpe Ratio stands at 0.52. The 5% lower confidence interval bound is **-0.01**, indicating marginal significance at the 95% level; while it technically touches zero, the vast majority of the probability mass is positive. In contrast, the canonical model's lower bound falls deep into negative territory (-0.31), suggesting its risk-adjusted performance is statistically indistinguishable from zero.



(a) canonical (Smart Decay) (b) MAT (Conviction)

Figure 9: **Bootstrap robustness analysis.** (a) The canonical model shows a lower mean Sharpe (0.27) and a confidence interval that widely encompasses zero (-0.31), indicating a lack of statistical significance. (b) The MAT model achieves a higher mean Sharpe (0.52) and a superior win rate (52.2%), with a confidence interval that is borderline significant (-0.01), reflecting a more robust predictive signal.

F.4 Section conclusion

The portfolio analysis confirms that the architectural innovations of the Modality-Aware Transformer translate into tangible improvements in downstream decision-making. By explicitly modeling the

interaction between text and time series, the MAT model achieves superior ranking capabilities, higher short-term precision, and a more robust separation between winners and losers compared to the canonical baseline.

However, it is crucial to contextualize these results within the efficient market hypothesis. While the MAT model demonstrates relative superiority, the absolute magnitude of the performance metrics remains modest. The Champion strategy achieves a gross Sharpe Ratio of 0.52, a figure that, while statistically significant, would likely be eroded by transaction costs and market impact in a live trading environment. This low absolute ceiling reflects the high efficiency of the S&P 500 universe, where alpha is notoriously difficult to extract using public data alone.

Ultimately, this analysis serves its primary purpose: it acts as a diagnostic tool validating that modality-aware attention successfully recovers predictive signals that are lost by standard fusion strategies, even if those signals require further refinement or broader data universes to become commercially viable.