Multimodal Language Models for Financial Forecasting from Interleaved Sequences of Text and Time Series

Ross Koval[1,3], Nicholas Andrews[2], and Xifeng Yan[1]

[1]University of California, Santa Barbara
[2]Johns Hopkins University
[3]AJO Vista
rkoval@ucsb.edu

Abstract

Text and time series data offer complementary views of financial markets: news articles provide narrative context about company events, while stock prices reflect how markets react to those events. However, despite their complementary nature, effectively integrating these interleaved modalities for improved forecasting remains challenging. In this work, we propose a unified neural architecture that models these interleaved sequences using modality-specific experts, allowing the model to learn unique time series patterns, while still enabling joint reasoning across modalities and preserving pretrained language understanding capabilities. To further improve multimodal understanding, we introduce a cross-modal alignment framework with a salient token weighting mechanism that learns to align representations across modalities with a focus on the most informative tokens. We demonstrate the effectiveness of our approach on a large-scale financial forecasting task, achieving state-of-the-art performance across a wide variety of strong unimodal and multimodal baselines. We develop an interpretability method that reveals insights into the value of time series-context and reinforces the design of our cross-modal alignment objective. Finally, we demonstrate that these improvements translate to meaningful economic gains in investment simulations.

1 Introduction

Text and time series provide complementary perspectives on financial markets. News articles describe company events, such as earnings announcements, product launches, and mergers, while stock prices reflect how markets react to these events over time. When these modalities are combined, it presents a promising yet challenging opportunity for large language models (LLMs) to enhance financial forecasting by reasoning over temporally aligned but semantically distinct inputs.
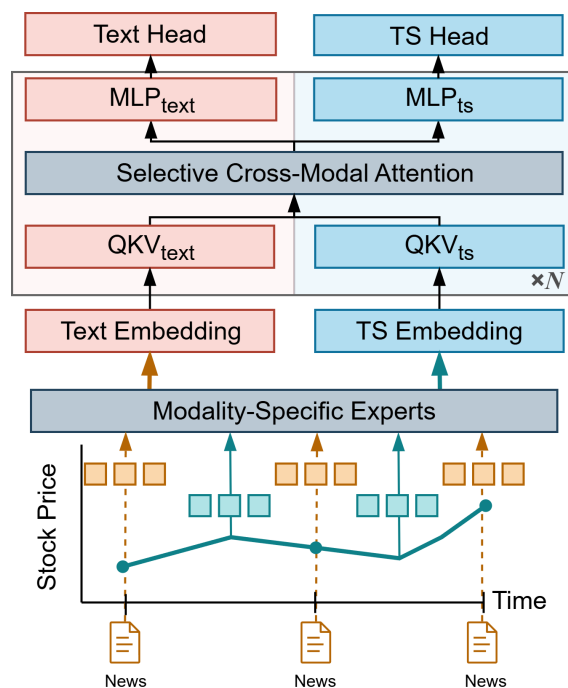

Figure 1: Overview of our multimodal forecasting task and proposed model (MSE-ITT), which processes interleaved sequences of tokens of news articles (Text) and discretized stock returns (TS). MSE-ITT incorporates modality-specific experts to capture distinct patterns in text and time series, while enabling joint reasoning across modalities through selective cross-modal attention.

To better understand their complementary nature, consider how time series and text contribute distinct but related signals. The time series data offers global context into both short and long-term price behavior, allowing the model to learn patterns driven by investor biases (Jegadeesh, 1990; Jegadeesh and Titman, 1993; Kelly et al., 2021). Moreover, the interleaved stock prices act as implicit supervision, revealing how markets have previously responded to similar news, and exposing the model to repeated cause-and-effect dynamics. Conversely, news articles provide both retrospec-

tive context about historical market behavior and forward-looking signals that may reinforce or contradict current price trends. By jointly modeling these modalities, LLMs have the potential to learn cross-modal interactions and context-aware representations that would not be possible from either modality alone.

However, effectively integrating long sequences of text and time series in a unified model remains challenging. Simple strategies, such as converting numerical data into strings of digits, fail to capture their distinct structures. Language is discrete, syntax-rich, and compositional, while time series are continuous, stochastic, and governed by temporal dependencies. Furthermore, news arrives at irregular intervals, while stock prices are observed daily. These structural differences make it difficult for pretrained LLMs, which are optimized for language, to extract meaningful signals from time series inputs. Instead, we argue that modeling these multimodal sequences requires modality-specific components that can respect and exploit the unique structure of each input type, while enabling joint reasoning across modalities.

To address this, we propose a unified multimodal architecture with modality-specific experts and design a pretraining objective to effectively learn cross-modal interactions. In summary, we make the following key contributions.

1. We design a multimodal architecture (**MSE-ITT**) that can effectively model interleaved sequences of text and time series, demonstrating state-of-the-art performance on a challenging financial forecasting task compared to a comprehensive set of baselines (§4, Table 2).

2. We introduce a unified cross-modal alignment framework (**SALMON**) with a dynamic, salient token weighting mechanism (**STW**) to effectively learn time-series-specific features that enhance language understanding (§4.2, Table 3).

3. We develop an interpretability method that reveals insights into the value of time series context and demonstrate that it translates to significant economic gains in investment applications (§7.1, Table 5, Table 6).

Broader Impact We hope this work encourages future research on LLMs that reason over interleaved sequences of text and time series, particularly in domains where structured and unstructured data interact over time, such as finance, healthcare, and climate. Our findings highlight the limitations of treating time series equivalently to language and underscore the importance of dedicated mechanisms for structured time series inputs. To support this research direction, we release the code and data at: `https://github.com/rosskoval/mlm_text_ts`.

2 Related Work

2.1 Multimodal Time Series Forecasting

LLMs have greatly improved their capabilities in understanding multimodal inputs, such as text, images, audio, and video (Liu et al., 2023a; Li et al., 2024; Team, 2024; Chen et al., 2023). However, they have demonstrated challenges in understanding time series data. While some work has found benefit in using LLMs to perform time series forecasting with zero-shot learning (Gruver et al., 2023) or finetuning (Jin et al., 2023), other work has found that their pretrained weights do not provide positive transfer (Tan et al., 2024) and struggle to reason about them effectively (Merrill et al., 2024) without specialized encoding (Chow et al., 2024). Moreover, recent work on contextualized forecasting has explored integrating text and time-series data using two main strategies. One paradigm converts time series into strings or embeddings, allowing text to condition LLM predictions (Jin et al., 2023; Williams et al., 2024; Kim et al., 2024b). However, this assumes LLMs can natively interpret time series and limits the model's ability to learn modality-specific patterns. Another approach uses frozen language models to extract fixed text features for multivariate time series models (Liu et al., 2024; Li et al., 2025). While this allows modality-specific patterns, it prevents deep cross-modal interaction and inhibits the reasoning capabilities of LLMs.

2.2 Financial Prediction

Recent work has shown that language models can predict stock returns from news articles (Chen et al., 2022; Xie et al., 2023; Lopez-Lira and Tang, 2023; Koval et al., 2025). Separately, historical price patterns have been shown to be predictive (Jegadeesh, 1990; Jegadeesh and Titman, 1993; Kelly et al., 2021). These findings have motivated recent work exploring methods to integrate textual and time series data to improve forecasting (Xu and Cohen, 2018; Ang and Lim, 2022; Koval et al., 2024; Wang

et al., 2024a; Zong et al., 2024; Mou et al., 2025). However, most of these models process each modality independently and generally rely on late-fusion methods, limiting cross-modal interaction during representation learning.

3 Problem Statement

3.1 Multimodal Inputs

In our main problem formulation, we consider a multimodal sequence of time-stamped inputs, aperiodically arriving news articles (text), and daily stock returns (time series), described below.

Time Series Inputs As time series inputs X_t, we consider the daily stock return r_t of the company over the last 1-year (252 trading days).

$$X_t = \{r_{t-252}, ..., r_{t-1}\}$$

While we demonstrate that our method generalizes to multivariate time series in §A.10, we focus on the univariate case in our main experiments to more precisely study the effects of the cross-modal interaction method in a controlled setting.

Textual Inputs As textual inputs Y_t, we consider the N most recent news articles a_t about the company prior to time t:

$$Y_t = \{a_{t-N}, ..., a_{t-1}\}$$

For computational efficiency, we select the $N = 10$ most recent articles about the same company that occurred within the last 1-year.

3.2 Task Formulation

Following Xu and Cohen (2018); Chen et al. (2022), we adopt the task of predicting the direction of the stock price P_t change over the course of short-term and long-term horizons (days) $h \in \{7D, 30D\}$ at prediction time t.

$$D_t = \text{Sign}(P_{t+h} - P_t)$$

We evaluate the performance on this binary classification task with AUC because the continuous scores are more informative than discrete classes for investment management applications.

Data Acquisition and Curation We curate financial news articles in English from the FNSPID Dataset (Dong et al., 2024), for US-based public companies, which covers a variety of company events and news sources. We perform a curation

process, following the filtering criterion in Chen et al. (2022) for data quality, detailed in §A.4. Our sample contains more than 3,000 public companies in the US, encompassing a diverse range of firm sizes and industries.

Data Statistics and Task Formulation We temporally partition the data into training (2010-2017), validation (2018-2019), and test (2020-2024) sets. We provide summary statistics in Table 1.

	Train	Validation	Test
Start Date	Jan-2010	Jan-2018	Jan-2020
End Date	Dec-2017	Dec-2019	Dec-2024
# Samples	155,146	36,931	115,611
# Companies	2,591	2,249	3,564

Table 1: Summary statistics of the characteristics of financial news articles and stock return time series in each sample split.

4 Proposed Method

4.1 Model Design

In this section, we introduce our proposed modality-specific experts multimodal model for interleaved sequences of text and time series (**MSE-ITT**), illustrated in Figure 1. The use of mixture-of-experts (MOE) layers (Shazeer et al., 2017; Fedus et al., 2022; Zoph et al., 2022) in LLMs have become pervasive because of their improved representational capacity and compute efficiency, allowing subnetworks to specialize in different regions of the input. Inspired by recent advances in multimodal MOE-based LMs for text and images (Zhou et al., 2024; Shi et al., 2024a; Lin et al., 2024; Liang et al., 2025), we design a modality-specific experts architecture that builds on Llama3-8B (Grattafiori and et al., 2024), an autoregressive LM with strong language capabilities, and extend it with components tailored for temporal and cross-modal understanding.

Modality-Specific Experts To process both modalities effectively, we introduce dedicated modality-specific (TS) parameters into each layer of the pretrained LM. These modality-specific components are responsible for processing time series inputs, enabling the model to capture time-series patterns without disrupting the pretrained language capabilities of the base LM. This separation reduces cross-modal interference and imposes an inductive bias that respects the structural differences

between text and time series. At the same time, it allows for joint modeling of interleaved sequences. We maintain causal masking in the self-attention layers to preserve temporal causality.

These added parameters include layer normalization (LN), multi-headed attention projections (QKV), and feedforward MLPs (MLP):

$$\mathbf{h}_{\text{text}}^q, \mathbf{h}_{\text{text}}^k, \mathbf{h}_{\text{text}}^v = \text{QKV}_{\text{text}}\big(\text{LN}_{\text{text}}(\mathbf{h}_{\text{text}})\big)$$

$$\mathbf{h}_{\text{ts}}^q, \mathbf{h}_{\text{ts}}^k, \mathbf{h}_{\text{ts}}^v = \text{QKV}_{\text{ts}}\big(\text{LN}_{\text{ts}}(\mathbf{h}_{\text{ts}})\big)$$

$$\mathbf{h}_{\text{text}} = \text{MLP}_{\text{text}}\big(\text{LN}_{\text{text}}(\mathbf{h}_{\text{text}})\big)$$

$$\mathbf{h}_{\text{ts}} = \text{MLP}_{\text{ts}}\big(\text{LN}_{\text{ts}}(\mathbf{h}_{\text{ts}})\big)$$

In this design, the text ($\mathbf{h}_{\text{text}}$) and time series ($\mathbf{h}_{\text{ts}}$) hidden states are routed to separate modules, before performing joint self-attention over the multimodal sequence:

$$\mathbf{h} = \text{softmax}\Big([\mathbf{h}_{\text{text}}^q; \mathbf{h}_{\text{ts}}^q][\mathbf{h}_{\text{text}}^k; \mathbf{h}_{\text{ts}}^k]^\top\Big)\,[\mathbf{h}_{\text{text}}^v; \mathbf{h}_{\text{ts}}^v]$$

We initialize these dedicated TS-parameters from the pretrained values in the corresponding LM layer. We omit the residual connections and scaling for brevity.

Selective Cross-Modal Attention Prior work has shown that early layers of pretrained LMs capture low-level, syntactic patterns (Clark et al., 2019; Nagrani et al., 2021; Zhang et al., 2025), while deeper layers encode higher-level, global semantics. In addition, attention weights have been found to exhibit noise due to positional biases (Liu et al., 2023b; Xiao et al., 2023; Ye et al., 2024; Golovneva et al., 2025). Based on these findings, our hypothesis is that deeper layers are more likely to benefit from time-series context and that introducing cross-modality attention at early layers may be harmful, as it risks disrupting low-level, modality-specific representations with noisy time-series signals. To address this, we restrict cross-modality attention to the top half of the network layers (16-32), allowing the lower layers (1-16) to focus on learning more localized, modality-specific features. This design choice results in efficiency gains, and, as shown in our ablation studies (Table 4), improves language understanding and task performance.

Interleaved Multimodal Sequence We design the input as a temporally ordered sequence of multimodal tokens, where a_t is a sequence of tokens

for news article a and r_t is the stock return with timestamp t.

$$x_{1:L} = \{r_{t-252}, ..., a_{t-N}, ..., r_{t-1}, ..., a_{t-1}\}$$

Since the news articles arrive at irregular intervals, we employ pointwise embedding tokenization (Shi et al., 2024b) to accommodate a variable number of time steps between news articles, rather than using fixed-length patches. This allows flexible modeling of sequences with variable temporal resolution and supports application to other domains with irregular event streams.

Because financial data is known to be statistically noisy, we discretize the continuous time series values into a fixed number of bins B, enabling more robust representations and reducing sensitivity to outliers. We learn embeddings (TSEmb) for each bin (Rabanser et al., 2020; Ansari et al., 2024) with quantile binning to ensure a well-calibrated distribution. We tune B over the validation set and ablate this design choice in §A.8. This choice allows us to transform the inputs into a unified sequence of embeddings z_i:

$$\mathbf{z}_i = \begin{cases} \text{TextEmb}(x_i), & \text{if } \text{mod}(x_i) = \text{text}, \\ \text{TSEmb}(x_i), & \text{if } \text{mod}(x_i) = \text{ts}, \end{cases}$$

We pass them through MSE-ITT to obtain contextualized hidden states $\mathbf{h}$:

$$\mathbf{h}_{1:L} = \text{MSE-ITT}(\mathbf{z}_{1:L}) = [\,\mathbf{h}_1, \dots, \mathbf{h}_L\,]$$

This interleaved input design enables the model to leverage the pretrained LLM's relative positional encodings (Su et al., 2024), to reflect the natural temporal order of events, rather than relying on learnable time embeddings (Woo et al., 2024) only at the input layer. Because rotary encodings impose a strong locality bias throughout the network, preserving temporal order in the input allows the model to better exploit this inductive bias in its attention patterns. It also ensures that the relative distance between article tokens reflects their actual separation in time, spaced by time series embeddings corresponding to the number of days between events.

4.2 SALMON

To further improve multimodal understanding, we introduce Salience-Aware Language Modeling over Interleaved Modalities (**SALMON**), a unified pretraining objective, that jointly predicts the next

text token and the next time series token in an interleaved multimodal sequence, with a dynamic salience-based weighting mechanism. Since we discretize the time series into discrete tokens, both objectives can be trained using cross-entropy loss, with separate projection heads for text tokens (U_{text}) and time series tokens (U_{ts}):

$$P_\theta(x_i) = \begin{cases} \text{softmax}(U_{\text{text}}\mathbf{h}_{i-1}), & \text{if mod}(x_i) = \text{text}, \\ \text{softmax}(U_{\text{ts}}\mathbf{h}_{i-1}), & \text{if mod}(x_i) = \text{ts}, \end{cases}$$

$$\mathcal{L}(x;\theta) = - \sum_{i \in \mathcal{I}_{\text{text}}} \log P_\theta(x_{i,\text{text}}) \; - \sum_{i \in \mathcal{I}_{\text{ts}}} \log P_\theta(x_{i,\text{ts}})$$

To enable cross-modal learning without disrupting pretrained language capabilities, we freeze the pretrained text-branch parameters and train only the newly introduced TS-specific parameters. This joint objective encourages the model to learn to identify features in the time series that are predictive of future news text, and learn aligned representations that capture how text and time series data co-evolve together.

Salient Token Weighting (STW) While SALMON enables joint language modeling across modalities, standard cross-entropy loss places equal weight on all next token predictions. However, we hypothesize that not all textual tokens should theoretically or equally benefit from time-series context. In typical news text, many tokens, such as function or filler words, are easily predicted using neighboring tokens. While these tokens may not benefit from TS-context, we suspect that specific salient tokens, such as those sentiment-charged or related to market behavior, could benefit substantially. For example, consider the news headline in Figure 2.

We hypothesize that the easily predicted words dominate the loss function and that a selective weighting mechanism can help focus the training signal on tokens where TS-context provides significant mutual information, thereby improving the alignment process by reducing weight on noisy, irrelevant tokens.

To systematically identify such salient tokens without relying on external sentiment dictionaries, we design a contrastive estimation approach inspired by recent advances in contrastive decoding (Li et al., 2023; Yuan et al., 2024) and long-context training (Ye et al., 2024; Fang et al., 2024). To this end, we leverage the LM with text-only inputs as
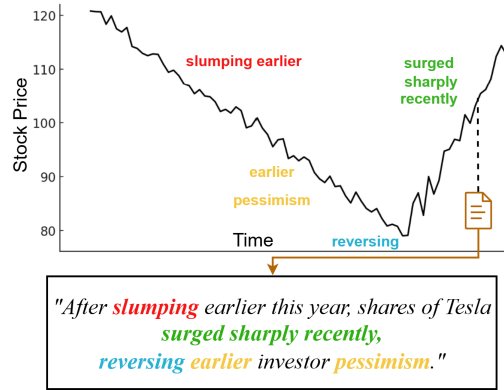

Figure 2: Illustration of our cross-modal alignment framework, SALMON, which learns to align historical stock price behavior and news articles with a unified objective. The Salient Token Weighting (STW) mechanism dynamically assigns higher weight to tokens that benefit most from time-series context, improving cross-modal alignment.

a contrastive baseline. Specifically, we compute two versions of token-level predicted probabilities: one using text-only inputs and another using the full multimodal input. Since we freeze the text-only parameters of the model, we can compute the baseline forward pass efficiently with no gradients attached. The ratio of these probabilities (i.e. the likelihood ratio) reflects how much the TS context improves the prediction of each token, and serves as a proxy for token salience.

Mathematically, assume that $P_\theta(x_{i,\text{text}} \mid x_{j<i,\text{text}})$ is the probability of the i-th text token given preceding text tokens, and $P_\theta(x_{i,\text{text}} \mid x_{j<i,\text{text}}, x_{j<i,\text{ts}})$ is the probability when preceding time series context is also provided. We define the token-level weight as:

$$W(x_{i,\text{text}}) = \frac{P_\theta(x_{i,\text{text}} \mid x_{j<i,\text{text}}, x_{j<i,\text{ts}})}{P_\theta(x_{i,\text{text}} \mid x_{j<i,\text{text}})}$$

such that W directly measures how much the prediction of each token improves when time series context is available. Values greater than 1 indicate tokens that benefit from time series context, while values less than 1 indicate tokens where time series context is less helpful or potentially distracting. Finally, the weights are normalized to have mean one per sequence $\tilde{W}$ and then applied to each textual token in the cross-entropy loss. Mathematically, let $x_{i,\text{text}}$ and $x_{i,\text{ts}}$ denote the i-th text and time series tokens, respectively, and $c_{j<i} = (x_{j<i,\text{text}}, x_{j<i,\text{ts}})$ denote the previous context, then the STW loss is

Model Class	Model	Base LM	Input	7D	30D
Zero-Shot LLM	GPT-4o, Direct (Lopez-Lira and Tang, 2023)	GPT-4o	text	52.45	53.31
	GPT-4o, CoT	GPT-4o	text	53.15	54.33
	GPT-4o, Direct	GPT-4o	ts	50.92	49.77
	GPT-4o, CoT	GPT-4o	ts	48.70	47.90
	GPT-4o, Direct (Williams et al., 2024)	GPT-4o	text, ts	52.09	52.71
	GPT-4o, CoT (Tan et al., 2025)	GPT-4o	text, ts	50.56	53.05
Unimodal	TS-Only (Nie et al., 2022)	None	ts	52.93	53.88
	Text-Only (Chen et al., 2022)	Llama3-8B	text	53.76	54.13
MMTSF	TimeLLM (Jin et al., 2023)	Llama2-7B	text, ts	53.79	55.10
	TaTs (Li et al., 2025)	Llama2-7B	text, ts	54.48	54.81
	TTSR (Chow et al., 2024)	Mistral-7B	text, ts	55.93	56.17**
	TimeMDD (Liu et al., 2024)	Llama3-8B	text, ts	55.15	55.25
	Hybrid-MMF (Kim et al., 2024b)	Llama3-8B	text, ts	55.96*	55.84
SFF	FinMA (Xie et al., 2023)	Llama2-7B	text, ts	51.11	52.15
	MTFE-MICM (Koval et al., 2024)	BigBird-125M	text, ts	55.44	54.49
	FININ (Wang et al., 2024a)	RoBERTa-125M	text, ts	52.47	53.13
	StockTime (Wang et al., 2024b)	Llama3-8B	text, ts	55.36	55.85
	MAT (Emami Gohari et al., 2024)	FinBERT-110M	text, ts	54.43	53.81
	MM-iTransformer (Mou et al., 2025)	FinBERT-110M	text, ts	54.16	53.57
Proposed	**MSE-ITT**	Llama3-8B	text, ts	**57.94*** [0.08]	**58.48*** [0.07]

Table 2: Main Results (higher is better): Model performance on the test set of our multimodal prediction task at different forecasting horizons. All results indicate the AUC of the model's predicted probabilities reported in percentage units. "[]" indicate the sample standard deviation of the results over 3 training runs with different random seeds. *, ** indicate that the performance of our proposed model is statistically better ($p < 0.01$) than the next best performing model according to DeLong's test. The last row indicates our proposed method MSE-ITT.

formed on textual inputs x_{text} by:

$$\mathcal{L}_{STW}(x_{\text{text}}; \theta) = - \sum_{i \in \mathcal{I}_{\text{text}}} \tilde{W}(x_{i,\text{text}}) \cdot$$
$$\log P_\theta(x_{i,\text{text}} \mid c_{j<i})$$

This weighting mechanism amplifies the learning signal for textual tokens that benefit most from time series context. Since, at the beginning of training, the estimated token-level weights are not meaningful as the model has not yet learned to effectively leverage the time series context, we warm-start $W = 1$ for the first 20% of training steps, and then relax this constraint to $W \in [0.1, 10.0]$ as the models learn to better utilize this information. We perform this cross-modal pretraining step on the input training data, and ablate the value of it and the token weighting mechanism in Table 3.

Implementation Details Our method is parameter and compute efficient as we freeze the pretrained text-branch parameters, and only finetune the newly added TS-branch parameters efficiently with LoRA (Hu et al., 2021), during both cross-modal alignment and task finetuning.

5 Baselines

We provide a comprehensive set of strong baselines to evaluate the benefits of our approach, spanning commercial LLMs, state-of-the-art multimodal time series forecasting models, and specialized financial models for stock movement prediction.

Unimodal First, we provide simple unimodal baselines that indicate the independent forecasting ability of each input modality. **Text-Only** indicates training a classifier on frozen Llama3-8B embeddings of the text-only inputs (Ke et al., 2019). **TS-Only** indicates training PatchTST (Nie et al., 2022) on only the time series inputs.

Zero-Shot LLMs We include the zero-shot performance of GPT-4o (OpenAI, 2024) with a variety of different prompting configurations, following (Williams et al., 2024), establishing state-of-the-art commercial LLM baselines and highlighting the difficulty of the task. We provide different prompting methods, including direct prediction (**Direct**) and Chain-of-Thought (**CoT**) (Kim et al., 2024a; Tan et al., 2025). The time series inputs are converted to text strings and we tune their formatting with validation set performance (Gruver et al., 2023; Williams et al., 2024). We also provide text-only and time-series only baselines to indicate the relative ability of LLMs to reason over each modality, detailed in §A.5.

992

Multimodal Time Series Baselines We implement a variety of baselines specialized for multimodal time series forecasting (**MMTSF**). These include TaTs (Li et al., 2025), TimeMDD (Liu et al., 2024), TimeLLM (Jin et al., 2023), Hybrid-MMF (Kim et al., 2024b), and TTSR (Chow et al., 2024). All of these models are finetuned on the training data according to their original implementations, described further in §A.2.

Financial Forecasting Baselines We include a variety of baselines specialized for multimodal stock movement prediction (**SFF**). These include FinMA (Xie et al., 2023), MTFE-MICM (Koval et al., 2024), FININ (Wang et al., 2024a), StockTime (Wang et al., 2024b), MAT (Emami Gohari et al., 2024), and MM-iTransformer (Mou et al., 2025). All of these models are finetuned on the training data, detailed in §A.2.

6 Experimental Results and Analysis

6.1 Main Results

Overall, we find that our proposed model delivers meaningful gains in forecasting performance across time horizons compared to a strong set of diverse baselines, and that these gains translate to significant improvements in investment simulations (Table 6). As we further demonstrate with ablation experiments (Table 3, Table 4), these improvements stem from key architectural design choices grounded in inductive biases. First, the model captures modality-specific structure by introducing dedicated parameters for time series data, while maintaining a unified architecture that enables joint reasoning across modalities. Second, our SALMON objective aligns time series features in the latent space of the LLM for enhanced cross-modal understanding, while our dynamic salient weighting mechanism enhances cross-modal alignment by focusing on key tokens that benefit most from time series context.

Time Series Context The main results in Table 2 highlight the challenging nature of the task and that the value of multimodal context highly depends upon the method of integration. While some methods benefit significantly from multimodal context, others underperform unimodal baselines.

We highlight that across both general and specialized baselines, there is a clear trend that methods that initially encode the time series with separate parameters from the LLM perform better than those that exclusively rely on LLM's native ability to extract meaningful time series features in either the text or latent space. Similarly, we find that models that simply treat text features as additional channels within a multivariate time series model underperform those that harness the powerful reasoning capabilities of LMs. These findings reinforce our design decision for modality-specific experts within a unified architecture.

Zero-Shot LLMs We highlight two interesting findings for GPT-4o. Firstly, GPT-4o performs better with text-only inputs than multimodal inputs, failing to benefit from the time series context. Secondly, prompting with chain-of-thought improves the ability of GPT-4o to analyze the textual inputs while failing to improve its ability to reason about the time series inputs. Overall, these results clearly highlight the challenges that LLMs exhibit in reasoning over time series inputs (Merrill et al., 2024), particularly in the financial domain, which are often noisy and lack consistent patterns.

Method	7D	30D
MSE-ITT w/o SALMON	56.93	57.14
+SALMON w/o STW	57.56*	57.89*
+SALMON w/ STW	57.94**	58.48**

Table 3: Results demonstrate the value of our proposed Cross-Modal alignment objective (SALMON) with Salient-Token Weighting (STW) compared to the baseline MSE-ITT model finetuned without any pretraining. *, ** indicate that the performance of the model is statistically better ($p < 0.05$) than the previous model according to DeLong's test.

6.2 Ablation Studies

We conduct extensive ablations of our multimodal architecture and cross-modal alignment process, highlighting the contribution of each design choice to overall model performance.

Cross-Modal Alignment In Table 3, we find that SALMON pretraining significantly improves both language understanding capabilities and task performance by aligning representations across modalities. Additionally, our STW mechanism provides additional benefits over standard equal token weighting by selectively identifying and emphasizing tokens that benefit most from TS context, leading to more effective cross-modal representation learning.

Model	LM Loss	30D
Text-Only	2.20	54.13
Shared Parameters	2.00	55.80
Separate QKV	1.85	56.78
Separate MLP	1.91	56.31
Cross-Modal Attention, Layers 1-16	1.96	56.00
Cross-Modal Attention, Layers 1-32	1.81	56.51
MSE-ITT	1.78	57.14

Table 4: Results demonstrate the value of our proposed MSE-ITT architecture in both language understanding (LM Loss) and financial forecasting (30D) performance, compared to two sets of baselines that (1) share parameters across text and TS inputs and (2) perform cross-modal attention in different layers of the network. Note that the LM Loss results include cross-modal alignment training (SALMON) but the 30D results do NOT.

Modality-Specific Experts   In Table 4, we provide a direct comparison of our method against baseline models that share weights between the text and time-series inputs. These results demonstrate that our multimodal model effectively leverages time series context to improve its language understanding capabilities with unified pretraining (20% reduction in LM loss). Compared to sharing parameters, our modality-specific experts architecture improves both language and time series understanding capability, mitigating cross-modal interference and allowing unique modality patterns.

Selective Cross-Modal Attention   We include baselines in which cross-modal attention is performed in early layers of the network, rather than just the second half. As shown in Table 4, cross-modal attention in the early layers is detrimental to performance, supporting our hypothesis that early LM layers focus on low-level features that can be disrupted by premature cross-modal interactions. We believe these findings could generalize to other multimodal contexts (e.g. healthcare, climate, etc.) in which modalities possess different properties but shared temporal alignment.

7   Model Analysis and Interpretability

In this section, we explore the behavior of the model and reveal insights into the value of time series context and its potential impact on investment applications.

7.1   Value of Time Series Context

To further investigate where and when time-series context is most beneficial, we use the LM Finan-

Word Category	Likelihood Ratio
Stop Words	0.71
Non-Sentiment	1.45
All-Sentiment	1.83
Positive	2.17
Negative	1.74
Litigious	1.76
Uncertainty	1.63
Constraining	1.23
Weak Modal	2.95
Strong Modal	1.28

Table 5: Median likelihood ratios across word categories from the LM Financial Dictionary (Loughran and McDonald, 2011), computed on the test set. These ratios quantify the marginal benefit of time series context for predicting words in each category, revealing the strongest gains for sentiment-charged words.

cial Dictionary (Loughran and McDonald, 2011) to classify the financial sentiment of words. Following §4.2, we compute the benefit of TS-context (likelihood ratio) for each group of words across the test set with our MSE-ITT model after SALMON pretraining.

In Table 5, we report median statistics on the likelihood ratio for sentiment words, as defined by the LM Financial Dictionary (Loughran and McDonald, 2011), which indicates the marginal benefit that time series context provides for predicting the identity of the tokens. These results demonstrate that sentiment-charged words benefit from the time series context significantly more than non-sentiment words, and dramatically more than stop words. These findings highlight the complementary nature of time series and textual data and reinforce the design of our **STW** loss function §4.2. The time series context provides valuable information for improving language understanding and interpreting the impact of news events.

7.2   Portfolio Simulations

In Table 6, we demonstrate the economic value of our methodology with portfolio simulations. We form monthly long-short (*market-neutral*) portfolios (Fama and French, 2015) by sorting stocks based on the 30D model predictions from the past month, detailed in §A.3. For comparison, we simulate unimodal baselines and the best performing multimodal baselines, described further in §A.3. Following Cong et al. (2021), we include net performance that includes conservative estimates of the impact of transaction costs on portfolio implementation.

The test period spans 5-years of highly diverse market regimes, including the COVID-19 crash

and recovery, stimulus-driven expansion, and inflation and rate hikes, ensuring robustness across economic conditions. The resulting performance of our unified multimodal model generates investment performance that is economically and statistically better than the best performing multimodal baselines. These results demonstrate that our model provides significant predictive value in a real-world trading setting.

Method	Net Return	Volatility	Net Sharpe Ratio
TS-Only (Nie et al., 2022)	5.99	13.11	0.46
Text-Only (Chen et al., 2022)	8.60	10.47	0.82
TTSR (Chow et al., 2024)	12.37	11.28	1.10
Hybrid-MMF (Kim et al., 2024b)	11.60	10.19	1.13
MTFE-MICM (Koval et al., 2024)	10.23	10.39	0.99
StockTime (Wang et al., 2024b)	13.91	12.65	1.10
Proposed, MSE-ITT	**17.01**	**11.26**	**1.51**

Table 6: Annualized portfolio statistics of simulated investment performance, expressed in percentage units. "Net" performance includes an estimate of the impact of transaction costs, detailed in §A.3.

8 Conclusion

We propose a unified multimodal architecture for modeling interleaved sequences of text and time series data, and introduce a cross-modal alignment framework with a salient token weighting mechanism that learns to align representations across modalities with a focus on the most informative tokens. Our approach demonstrates state-of-the-art performance on a challenging financial forecasting task, and our ablation experiments confirm the contribution of our design decisions. These forecasting improvements translate to economically meaningful gains in portfolio simulations, underscoring the real-world value of our approach. These findings highlight the need for modality-specific structures and joint reasoning in multimodal LMs, with broader implications for domains in which text and time series co-evolve.

Limitations Our findings demonstrate that large language models can benefit significantly from structured time series context when modeling interleaved sequences of text and numerical data, but only with the appropriate inductive biases. The proposed architecture improves both forecasting accuracy and simulated investment performance through modality-specific routing and specialized cross-modal pretraining. However, our experiments focus on financial forecasting within English financial news and US-based stocks, but we believe the methodology is applicable to other domains involving interleaved text and time series (e.g. healthcare, climate), which we leave to future work. While we incorporate conservative estimates of transaction costs in our investment simulations, real-world trading requires more detailed consideration of trade execution and risk management, which we leave to future work. Please note that our financial prediction system is intended for research use and that portfolio results are presented for illustration purposes only, not as investment advice.

References

Gary Ang and Ee-Peng Lim. 2022. Guided attention multimodal multitask financial forecasting with inter-company relationships and global and local news. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6313–6326.

Abdul Fatir Ansari, Lorenzo Stella, Caner Turkmen, Xiyuan Zhang, Pedro Mercado, Huibin Shen, Oleksandr Shchur, Syama Sundar Rangapuram, Sebastian Pineda Arango, Shubham Kapoor, et al. 2024. Chronos: Learning the language of time series. *arXiv preprint arXiv:2403.07815*.

Dogu Araci. 2019. Finbert: Financial sentiment analysis with pre-trained language models. *arXiv preprint arXiv:1908.10063*.

Avi Caciularu, Arman Cohan, Iz Beltagy, Matthew E Peters, Arie Cattan, and Ido Dagan. 2021. Cdlm: Cross-document language modeling. *arXiv preprint arXiv:2101.00406*.

Guo Chen, Yin-Dong Zheng, Jiahao Wang, Jilan Xu, Yifei Huang, Junting Pan, Yi Wang, Yali Wang, Yu Qiao, Tong Lu, et al. 2023. Videollm: Modeling video sequence with large language models. *arXiv preprint arXiv:2305.13292*.

Yifei Chen, Bryan T Kelly, and Dacheng Xiu. 2022. Expected returns and large language models. *Available at SSRN 4416687*.

Winnie Chow, Lauren Gardiner, Haraldur T. Hallgrímsson, Maxwell A. Xu, and Shirley You Ren. 2024. Towards Time Series Reasoning with LLMs. arXiv preprint arXiv:2409.11376. *Preprint*, arXiv:2409.11376. Presented at the NeurIPS 2024 Workshop on Time Series in the Age of Large Models.

Kevin Clark, Urvashi Khandelwal, Omer Levy, and Christopher D. Manning. 2019. What does BERT look at? an analysis of BERT's attention. In *Proceedings of the 2019 ACL Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 276–286, Florence, Italy. Association for Computational Linguistics.

Lin William Cong, Ke Tang, Jingyuan Wang, and Yang Zhang. 2021. Alphaportfolio: Direct construction through deep reinforcement learning and interpretable ai. *Available at SSRN 3554486*.

Elizabeth R DeLong, David M DeLong, and Daniel L Clarke-Pearson. 1988. Comparing the areas under two or more correlated receiver operating characteristic curves: a nonparametric approach. *Biometrics*, pages 837–845.

Zihan Dong, Xinyu Fan, and Zhiyuan Peng. 2024. FN-SPID: A Comprehensive Financial News Dataset in Time Series. arXiv preprint arXiv:2402.06698. *Preprint*, arXiv:2402.06698.

Hajar Emami Gohari, Xuan-Hong Dang, Syed Yousaf Shah, and Petros Zerfos. 2024. Modality-aware transformer for financial time series forecasting. In *Proceedings of the 5th ACM International Conference on AI in Finance*, pages 677–685.

Eugene F Fama and Kenneth R French. 2015. A five-factor asset pricing model. *Journal of financial economics*, 116(1):1–22.

Lizhe Fang, Yifei Wang, Zhaoyang Liu, Chenheng Zhang, Stefanie Jegelka, Jinyang Gao, Bolin Ding, and Yisen Wang. 2024. What is wrong with perplexity for long-context language modeling? *arXiv preprint arXiv:2410.23771*.

William Fedus, Barret Zoph, and Noam Shazeer. 2022. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39.

Olga Golovneva, Tianlu Wang, Jason Weston, and Sainbayar Sukhbaatar. 2025. Multi-token attention. *arXiv preprint arXiv:2504.00927*.

Aaron Grattafiori and et al. 2024. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

Nate Gruver, Marc Finzi, Shikai Qiu, and Andrew G Wilson. 2023. Large language models are zero-shot time series forecasters. *Advances in Neural Information Processing Systems*, 36:19622–19635.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *arXiv preprint arXiv:2106.09685*.

Narasimhan Jegadeesh. 1990. Evidence of predictable behavior of security returns. *The Journal of Finance*, 45(3):881–898.

Narasimhan Jegadeesh and Sheridan Titman. 1993. Returns to buying winners and selling losers: Implications for stock market efficiency. *The Journal of Finance*, 48(1):65–91.

Ming Jin, Shiyu Wang, Lintao Ma, Zhixuan Chu, James Y Zhang, Xiaoming Shi, Pin-Yu Chen, Yuxuan Liang, Yuan-Fang Li, Shirui Pan, et al. 2023. Time-llm: Time series forecasting by reprogramming large language models. *arXiv preprint arXiv:2310.01728*.

Zheng Tracy Ke, Bryan T Kelly, and Dacheng Xiu. 2019. Predicting returns with text data. Technical report, National Bureau of Economic Research.

Bryan T. Kelly, Tobias J. Moskowitz, and Seth Pruitt. 2021. Understanding momentum and reversal. *Journal of Financial Economics*, 140(3):726–743.

Alex Kim, Maximilian Muhn, and Valeri Nikolaev. 2024a. Financial statement analysis with large language models. *arXiv preprint arXiv:2407.17866*.

Kai Kim, Howard Tsai, Rajat Sen, Abhimanyu Das, Zihao Zhou, Abhishek Tanpure, Mathew Luo, and Rose Yu. 2024b. Multi-Modal Forecaster: Jointly Predicting Time Series and Textual Data. arXiv preprint arXiv:2411.06735. *Preprint*, arXiv:2411.06735.

Ross Koval, Nicholas Andrews, and Xifeng Yan. 2024. Financial forecasting from textual and tabular time series. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 8289–8300, Miami, Florida, USA. Association for Computational Linguistics.

Ross Koval, Nicholas Andrews, and Xifeng Yan. 2025. Context-aware language models for forecasting market impact from sequences of financial news. *arXiv preprint arXiv:2509.12519*.

Dongting Li, Chenchong Tang, and Han Liu. 2024. Audio-llm: Activating the capabilities of large language models to comprehend audio data. In *International Symposium on Neural Networks*, pages 133–142. Springer.

Xiang Lisa Li, Ari Holtzman, Daniel Fried, Percy Liang, Jason Eisner, Tatsunori Hashimoto, Luke Zettlemoyer, and Mike Lewis. 2023. Contrastive decoding: Open-ended text generation as optimization. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 12286–12312, Toronto, Canada. Association for Computational Linguistics.

Zihao Li, Xuanye Lin, Zeyu Liu, Jiaru Zou, Zhenyu Wu, Ling Zheng, Dawei Fu, Yan Zhu, Hendrik Hamann, Hanghang Tong, et al. 2025. Language in the Flow of Time: Time-Series-Paired Texts Weaved into a Unified Temporal Narrative. arXiv preprint arXiv:2502.08942. *Preprint*, arXiv:2502.08942.

Weixin Liang, Junhong Shen, Genghan Zhang, Ning Dong, Luke Zettlemoyer, and Lili Yu. 2025. Mixture-of-Mamba: Enhancing Multi-Modal State-Space Models with Modality-Aware Sparsity. arXiv preprint arXiv:2501.16295. *Preprint*, arXiv:2501.16295.

Xi Victoria Lin, Akshat Shrivastava, Liang Luo, Srinivasan Iyer, Mike Lewis, Gargi Ghosh, Luke Zettlemoyer, and Armen Aghajanyan. 2024. MoMa: Efficient Early-Fusion Pre-training with Mixture of Modality-Aware Experts. *Preprint*, arXiv:2407.21770.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023a. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916.

Haoxin Liu, Shangqing Xu, Zhiyuan Zhao, Lingkai Kong, Harshavardhan Kamarthi, Aditya B. Sasanur, Megha Sharma, Jiaming Cui, Qingsong Wen, Chao Zhang, and B. Aditya Prakash. 2024. Time-MMD: Multi-Domain Multimodal Dataset for Time Series Analysis. In *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, December 10-16, 2024, Vancouver, BC, Canada, Track on Datasets and Benchmarks*.

Nelson F Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2023b. Lost in the middle: How language models use long contexts. *arXiv preprint arXiv:2307.03172*.

Alejandro Lopez-Lira and Yuehua Tang. 2023. Can chatgpt forecast stock price movements? return predictability and large language models. *arXiv preprint arXiv:2304.07619*.

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

Mike A. Merrill, Mingtian Tan, Vinayak Gupta, Tom Hartvigsen, and Tim Althoff. 2024. Language Models Still Struggle to Zero-shot Reason about Time Series. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3512–3533, Miami, Florida, USA. Association for Computational Linguistics.

Shangyang Mou, Qiang Xue, Jinhui Chen, Tetsuya Takiguchi, and Yasuo Ariki. 2025. Mm-itransformer: A multimodal approach to economic time series forecasting with textual data. *Applied Sciences*, 15(3):1241.

Arsha Nagrani, Shan Yang, Anurag Arnab, Aren Jansen, Cordelia Schmid, and Chen Sun. 2021. Attention bottlenecks for multimodal fusion. In *NeurIPS*.

Yuqi Nie, Nam H Nguyen, Phanwadee Sinthong, and Jayant Kalagnanam. 2022. A time series is worth 64 words: Long-term forecasting with transformers. *arXiv preprint arXiv:2211.14730*.

OpenAI. 2024. Gpt-4o: Openai's new omnimodal model. https://openai.com/index/gpt-4o. Accessed: 2025-04-23.

Stephan Rabanser, Tim Januschowski, Valentin Flunkert, David Salinas, and Jan Gasthaus. 2020. The effectiveness of discretization in forecasting: An empirical study on neural time series models. *arXiv preprint arXiv:2005.10111*.

Noam Shazeer, Azalia Mirhoseini, Krzysztof Maziarz, Andy Davis, Quoc Le, Geoffrey Hinton, and Jeff Dean. 2017. Outrageously large neural networks: The sparsely-gated mixture-of-experts layer. *arXiv preprint arXiv:1701.06538*.

Weijia Shi, Xiaochuang Han, Chunting Zhou, Weixin Liang, Xi Victoria Lin, Luke Zettlemoyer, and Lili Yu. 2024a. LMFusion: Adapting Pretrained Language Models for Multimodal Generation. *Preprint*, arXiv:2412.15188.

Xiaoming Shi, Shiyu Wang, Yuqi Nie, Dianqi Li, Zhou Ye, Qingsong Wen, and Ming Jin. 2024b. Time-moe: Billion-scale time series foundation models with mixture of experts. *arXiv preprint arXiv:2409.16040*.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. 2024. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063.

Alexander Swade, Matthias X Hanauer, Harald Lohre, and David Blitz. 2023. Factor zoo (. zip). *Available at SSRN*.

Mingtian Tan, Mike A. Merrill, Zachary Gottesman, Tim Althoff, David Evans, and Tom Hartvigsen. 2025. Inferring Event Descriptions from Time Series with Language Models. arXiv preprint arXiv:2503.14190. *Preprint*, arXiv:2503.14190.

Mingtian Tan, Mike A. Merrill, Vinayak Gupta, Tim Althoff, and Tom Hartvigsen. 2024. Are Language Models Actually Useful for Time Series Forecasting? In *Advances in Neural Information Processing Systems 37*, pages 60162–60191. Curran Associates, Inc.

Chameleon Team. 2024. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*.

Mengyu Wang, Shay B. Cohen, and Tiejun Ma. 2024a. Modeling News Interactions and Influence for Financial Market Prediction. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3302–3314, Miami, Florida, USA. Association for Computational Linguistics.

Shengkun Wang, Taoran Ji, Linhan Wang, Yanshen Sun, Shang-Ching Liu, Amit Kumar, and Chang-Tien Lu. 2024b. Stocktime: A time series specialized large language model architecture for stock price prediction. *arXiv preprint arXiv:2409.08281*.

Andrew Robert Williams, Arjun Ashok, Étienne Marcotte, Valentina Zantedeschi, Jithendaraa Subramanian, Roland Riachi, James Requeima, Alexandre

997

Lacoste, Irina Rish, Nicolas Chapados, and Alexandre Drouin. 2024. Context is Key: A Benchmark for Forecasting with Essential Textual Information. arXiv preprint arXiv:2410.18959. *Preprint*, arXiv:2410.18959.

Gerald Woo, Chenghao Liu, Akshat Kumar, Caiming Xiong, Silvio Savarese, and Doyen Sahoo. 2024. Unified training of universal time series forecasting transformers.

Guangxuan Xiao, Yuandong Tian, Beidi Chen, Song Han, and Mike Lewis. 2023. Efficient streaming language models with attention sinks. *arXiv preprint arXiv:2309.17453*.

Qianqian Xie, Weiguang Han, Xiao Zhang, Yanzhao Lai, Min Peng, Alejandro Lopez-Lira, and Jimin Huang. 2023. Pixiu: A large language model, instruction data and evaluation benchmark for finance. *arXiv preprint arXiv:2306.05443*.

Yumo Xu and Shay B Cohen. 2018. Stock movement prediction from tweets and historical prices. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1970–1979.

Tianzhu Ye, Li Dong, Yuqing Xia, Yutao Sun, Yi Zhu, Gao Huang, and Furu Wei. 2024. Differential transformer. *arXiv preprint arXiv:2410.05258*.

Hongyi Yuan, Keming Lu, Fei Huang, Zheng Yuan, and Chang Zhou. 2024. Speculative contrastive decoding. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 56–64, Bangkok, Thailand. Association for Computational Linguistics.

Cheng Zhang, Jinxin Lv, Jingxu Cao, Jiachuan Sheng, Dawei Song, and Tiancheng Zhang. 2025. Unravelling the semantic mysteries of transformers layer by layer. *The Computer Journal*, page bxaf034.

Chunting Zhou, Lili Yu, Arun Babu, Kushal Tirumala, Michihiro Yasunaga, Leonid Shamis, Jacob Kahn, Xuezhe Ma, Luke Zettlemoyer, and Omer Levy. 2024. Transfusion: Predict the Next Token and Diffuse Images with One Multi-Modal Model. *Preprint*, arXiv:2408.11039.

Chang Zong, Jian Shao, Weiming Lu, and Yueting Zhuang. 2024. Stock movement prediction with multimodal stable fusion via gated cross-attention mechanism. *arXiv preprint arXiv:2406.06594*.

Barret Zoph, Irwan Bello, Sameer Kumar, Nan Du, Yanping Huang, Jeff Dean, Noam Shazeer, and William Fedus. 2022. St-moe: Designing stable and transferable sparse expert models. *arXiv preprint arXiv:2202.08906*.

A Appendix

A.1 Pretrained Language Models

We implement all models in PyTorch and source all pretrained checkpoints from HuggingFace.

A.2 Baseline Models

For our proposed model (MSE-ITT), for supervised classification, we train a classification head on top of the end-of-sequence token's last hidden representations to make binary predictions.

For some baseline models in which additional inputs or modalities are incorporated, such as proprietary sentiment scores (Wang et al., 2024a), we do not include these additional inputs (and we do not have access to them) in the model in order to present a fair comparison of model types and isolate the effects of multimodal fusion between sequences of text and time series. Some baseline models (Wang et al., 2024a; Liu et al., 2024; Li et al., 2025) explore the use of various LLMs for textual encoding. In such cases, we select the LM with the best reported performance for evaluation. For models that patch the time series input into non-overlapping consecutive chunks (Chow et al., 2024; Wang et al., 2024b), if the patch length used is not provided in the implementation details, then we tune the value over $\{1, 5, 10\}$ based on validation set performance. Further, for some multivariate time series baseline models that require a one-to-one mapping between text and time series inputs across time steps (Koval et al., 2024; Li et al., 2025), if there are no news articles on a given time step, then we simply carry forward the previous news article (embeddings) until the next news article (embeddings) are available. Additionally, in the MAT baseline (Emami Gohari et al., 2024), the authors do not report the number of topics or the base pretrained LM used in BERTTopic (Grootendorst, 2022), so we resort to using the hidden representations produced from their sentiment model FinBERT (Araci, 2019) as text features. Further, for the FinMA (Xie et al., 2023) baseline, we provide the text and time series sequences in the same format as for GPT-4o §A.5, such that the time series is converted to raw text and interleaved between news articles in temporal order.

For our proposed model, we use learnable special tokens to denote the beginning and end of each news article and modality (Caciularu et al., 2021), and include the article timestamp in the article text.

A.3 Portfolio Simulations

In Table 6, we demonstrate the economic value of our model predictions using portfolio simulations.

To perform these simulations in a realistic setting, we first filter the investment universe of stocks

according to sufficient liquidity requirements to ensure feasibility of portfolio implementation, including a minimum market capitalization of $250M and daily average value of shares traded of $1M.

Then, we form monthly long-short (market-neutral) quintile portfolios according to Fama and French (2015). To this end, we sort stocks based on the average 30D model predictions from news articles about companies in the past 1 month. Then our portfolios are formed by buying those in the top 20% of average scores and shorting those in the bottom 20% of average scores on a monthly basis in equal proportions. Please note that these portfolios are market-neutral and therefore have essentially no correlation with broad market indices.

In Table 6, we include conservative estimates of the impact of transactions costs on portfolio implementation. We follow the turnover-based method used in Cong et al. (2021), which conservatively estimates the annual transaction cost as 0.01 times the annual 1-way portfolio turnover. Therefore, the net return of the portfolio is the gross return minus the estimated transaction costs.

A.4 Data Curation

The FNSPID dataset spans multiple news sources, including Nasdaq, Reuters, CNBC, Benzinga, and cover a variety of company events, including product launches, earnings reports, and mergers. We truncate each article after the first 128 words during all experiments for computational efficiency. We only include articles originally written in English according to the following criteria (Chen et al., 2022): they are tagged relevant for only one company; they are longer than 100 characters or shorter than 10,000 characters; they contain less than 10% of numerical characters; they have less than 90% Jaccard similarity to a previous article (to remove potential duplicate articles). We require that each company possess at least 5 news articles within the past year as well as available stock returns to be in our sample. In addition, we apply further quality filtering to ensure that the dataset contains high-quality, event-driven news articles that are specific to individual stocks. We systematically remove broad market summaries, sector-level reports, and generic financial commentary by filtering out headlines containing specific keywords and patterns. This process allows us to isolate stock-specific events. There is a slight class imbalance so we randomly downsample the majority class to ensure balanced class ratios.

A.5 Zero-Shot LLMs

For the zero-shot GPT-4o baselines, we have explored a variety of prompting strategies, listed below. For this set of baselines, the time series inputs are converted to text strings and we tune their form {decimal, percentage}, digits of precision {2, 3, 4}, and whether to include a space in between digits, with validation set performance (Gruver et al., 2023; Williams et al., 2024). We do not find much variation in the performance of each, but tune them according to the validation set and report those results in Table 2.

To this end, we construct a natural language prompt that interleaves company-specific financial news with daily stock return sequences in chronological order. The model is tasked with predicting future price movement on a bounded scale.

Each prompt follows the format:

```
Given the company's daily stock
return over the last year
interleaved in temporal order
with recent news about the
company, sorted chronologically:
{interleaved_input}
Predict the company's future
stock price movement over the
next {horizon} days, on a
scale from 0 to 100, where
100 indicates strong positive
movement and 0 indicates
strong negative movement.
{prompt_style}
Prediction:
```

The field {interleaved_input} contains alternating sequences of news articles and their corresponding stock returns (e.g., News Article: ... Stock Returns: ...). The field {prompt_style} controls the reasoning strategy used by the model and takes one of two forms:

- **Direct**: The model immediately outputs a prediction.

- **Chain-of-Thought (CoT)**: The model is instructed to "think step-by-step" before producing a prediction, following recent findings that CoT improves LLM reasoning over time series data (Tan et al., 2025).

This design enables a uniform comparison of reasoning strategies across interleaved multimodal inputs, without any task-specific fine-tuning.

A.6 Statistical Significance

In Table 2, we report the sample standard deviation of results for our proposed method across 3 different training runs with different random seeds. The variability of results across random seeds stems from the randomness in the training process caused by random initialization of the classification layer weights and the random batch order of training samples during stochastic gradient descent optimization. We also report the results from DeLong's pairwise test of statistical significance between model AUC scores (DeLong et al., 1988).

A.7 Real-World Case Study

To further illustrate how our model benefits from multimodal reasoning, we examine a real-world event sequence involving the proposed Pfizer, Allergan merger, shown in Figure 3. This event unfolded over several months in 2015 and 2016, with headlines describing early merger rumors, confirmation of deal terms, political opposition, and ultimately, the termination of the transaction.

While a headline such as "Merger terminated by mutual consent" may appear negative in isolation, it follows a series of prior developments in which investors reacted negatively to news suggesting the deal would proceed. The accompanying stock return history reveals this clearly: mild declines on initial merger rumors, followed by a significant sell-off once the deal was announced, and a strong rebound when the US Treasury issued rules to block the deal. On April 6, 2016, when Pfizer and Allergan officially terminated the merger, the stock surged. This price behavior indicates that investors viewed the termination as favorable, even though the text of the headline appears quite negative.

To assess model behavior in this widely publicized and high impact event setting, we specifically train the text-only and ts-only baselines, and the proposed multimodal model (MSE-ITT) on point-in-time data and evaluate them on this example input. A text-only model, seeing the phrase "merger terminated", predicts a low probability (0.32) of a positive return. The time series-only model, reacting only to the long-term price trend of the stock price, driven negative by the merger rumors, predicts a score of (0.21). In contrast, our multimodal model (MSE-ITT), which jointly reasons over the headline, prior news, and accompanying stock price movements, assigns a high probability (0.79) to a positive return following the announce-

Date	News Headline	Stock Return	Interpretation
29 Oct 2015	Pfizer approaches Allergan about record merger	-1.9%, ... , -0.00%	Rumor; investors mildly optimistic
23 Nov 2015	$160 B deal announced; Pfizer to relocate to Ireland	–2.6%, ... , +2.3%	Fear of dilution and tax-inversion risk
5 Apr 2016	U.S. Treasury issues rules likely to scuttle tax inversion	+2.1%	Hope deal will be abandoned
6 Apr 2016	Merger **terminated** by mutual consent	+5.0%	Relief rally

Figure 3: Example sequence of news events and market reactions about the potential Pfizer-Allergan merger. The news articles contain persuasive language, yet the accompanying stock returns reveal the market's true perception: optimism on early rumors, a sharp sell-off once the deal terms are set, and a relief rally when the takeover is cancelled. Jointly modeling these inputs can more effectively interpret the outcome of such events.

ment. This prediction reflects the model's ability to perform a form of counterfactual reasoning: although the headline ("merger terminated") appears negative in isolation, the model recognizes that it cancels a deal investors had previously responded to unfavorably. By conditioning on prior market reactions to similar events, MSE-ITT infers that the termination is likely to be viewed positively. This deep integration of modalities enables the model to capture temporally grounded event semantics, where the sentiment of a headline depends critically on historical context.

Furthermore, this case study illustrates how unimodal models, lacking either narrative or market information, misinterpret semantically ambiguous events. In contrast, MSE-ITT aligns text and time series through token-level weighting that links financial language to observed outcomes, allowing it to resolve sentiment ambiguity and deliver more accurate, context-aware forecasts.

A.8 Time Series Discretization

For time-series discretization, we tune the number of discrete time-series bins $B \in \{4, 8, 16, 32, 64\}$ and their embedding dimension $d_{ts} \in \{32, 64, 128, 256, 512\}$. To map the time-series embeddings into the token embedding space of the LM, we learn a simple linear map to increase dimensionality from d_{ts} to d_{text}.

In Table 7, we compare our discretization approach to alternative continuous embedding methods. While the benefits of discretization are indeed modest and clearly not the primary source of our performance gains, they dramatically simplify and unify the interleaved multimodal (SALMON) pretraining objective. By converting all tokens, regard-

less of modality, to a common discrete space, we can apply a consistent cross-entropy loss to all discrete tokens, without inducing a complex multitask setup that requires careful calibration of continuous and discrete loss terms. In the continuous setting, we tune the weight on the TS (Mean Squared Error) loss over $w_{ts} \in \{0, 10, 100, 1000\}$ according to validation set performance.

Method	LM Loss	7D	30D
Discrete	1.78	57.94	58.48
Linear	1.83	57.54	58.10
MLP	1.83	57.62	58.13

Table 7: Results demonstrate the value of time series discretization in our multimodal architecture in both language understanding (LM Loss) and financial forecasting (7D, 30D) performance, compared to baselines that use continuous embeddings via linear and nonlinear transformations. These results include SALMON pretraining.

A.9 Training Details and Hyperparameter Tuning

We perform all experiments on a single NVIDIA H100 GPU with 80G in memory. We use AdamW to optimize all parameters. For all finetuned models, we use an effective batch size of 64 with gradient accumulation. We train all models for up to 5 epochs based on validation set performance for test evaluation. All supervised models are optimized using Binary Cross-Entropy as the loss function.

We tune the learning rate over {1e-6, 3e-6, 5e-6, 7e-6, 1e-5} and LoRA (Hu et al., 2021) rank parameter over $r \in \{16, 32, 64\}$ according to validation set performance. We apply LoRA adapters to all linear layers. We use a simple 2-layer MLP classification layer to project the token hidden states to make a binary prediction for all language models used.

For computational constraints, we train all models using mixed precision training and gradient checkpointing to satisfy GPU memory constraints, and clip gradient norms. For Llama-based models, we finetune in BF16 precision. For LoRA-based finetuning, we always set the value of the alpha parameter to be equal to double the value of rank parameter.

A.10 Multivariate Time Series

In this section, we demonstrate that our MSE-ITT model architecture can support numerical time series with multiple channels.

To this end, we select 15 commonly used market price and accounting-based financial variables available at the time of the report date from the definitions and cluster classifications in Swade et al. (2023). This set includes dividend yield (Value), earnings-to-price (Value), sales-to-price (Value), book value-to-price (Value), sales growth (Growth), earnings growth (Growth), gross profit to assets (Profitability), net income to equity (Profitability), net income to assets (Profitability), medium-term price momentum (Momentum), short-term price reversal (Reversal), price volatility (Low Risk), market leverage (Debt Issuance), share turnover (Low Risk), and market capitalization (Size).

Since some variables have different frequencies, ranging from daily to quarterly, we up-sample them all to daily frequency by forward-filling previous values.

We fit different discretized embeddings for each channel following our approach in §4. After this, we simply concatenate their embeddings together vertically at a daily frequency and similarly interleave with the news articles according to timestamp. To extend our SALMON pretraining objective to the multivariate setting, we learn a separate output projection head for each channel and compute the discretized token for each channel independently and average the loss across channels.

Method	LM Loss	30D
Univariate	–	57.14
Univariate w/ SALMON	1.78	58.48
Multivariate	–	58.45
Multivariate w/ SALMON	1.74	59.89

Table 8: Results indicate the performance of our MSE-ITT multimodal architecture when extended to the multivariate setting in both language understanding (LM Loss) and financial forecasting (30D) performance.