Financial Forecasting from Textual and Tabular 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

There is a variety of multimodal data pertinent to public companies, spanning from accounting statements, macroeconomic statistics, earnings conference calls, and financial reports. These diverse modalities capture the state of firms from a variety of different perspectives but requires complex interactions to reconcile in the formation of accurate financial predictions. The commonality between these different modalities is that they all represent a time series, typically observed for a particular firm at a quarterly horizon, providing the ability to model trends and variations of company data over time. However, the time series of these diverse modalities contains varying temporal and cross-channel patterns that are challenging to model without the appropriate inductive biases. In this work, we design a novel multimodal time series prediction task that includes numerical financial results, macroeconomic states, and long financial documents to predict next quarter's company earnings relative to analyst expectations. We explore a variety of approaches for this novel setting, establish strong unimodal baselines, and propose a multimodal model that exhibits state-of-the-art performance on this unique task. We demonstrate that each modality contains unique information and that the best performing model requires careful fusion of the different modalities in a multi-stage training approach. To better understand model behavior, we conduct a variety of probing experiments, reveal insights into the value of different modalities, and demonstrate the practical utility of our proposed method in a simulated trading setting.

1 Introduction

Investors are faced with the consumption of a myriad of diverse datasets relevant to public companies, spanning modalities, genres, and sources. In general, this data is released on a quarterly basis, and includes accounting statements, financial reports,

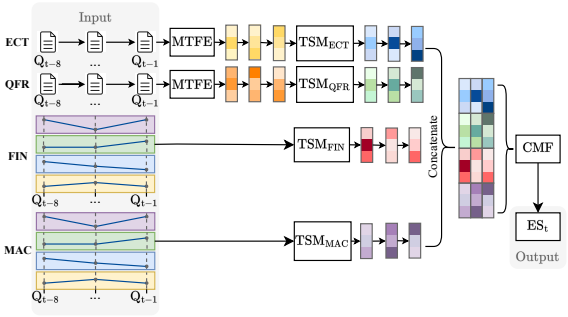

Figure 1: Overview of our multimodal time series prediction task and proposed multi-stage, modality-adaptive encoder fusion method. The inputs consist of the following time series: **AR** = autoregressive earnings surprise lags, **FIN** = tabular financial variables, **ECT** = earnings call transcripts, **QFR** = quarterly financial reports, and **MAC** = macroeconomic variables. These inputs are processed by the following model components: **TSM** = time series encoder, **MTFE** = multi-stage textual feature extraction, **CMF** = cross-modality fusion. The output is: **ES** = predicted Earnings Surprise (§3.1).

and earnings calls that comprehensively detail the current operations of the firm, recent financial performance, and discuss future business prospects and risks. While the accounting statements typically receive the most attention from investors, the textual content provided in earnings conference calls and financial reports is of equal importance as it reflects a direct communication between company executives and shareholders (Brown et al., 2004). This textual content provides two important qualitative sources of information relative to the financial metrics. First, it provides management context about how to interpret the current and historical financial results. Second, it also allows management to express their views about the future prospects of the company, providing the opportunity to capture the tone and sentiment from the most well-informed stakeholders. For instance, during periods of macroeconomic shocks, such as COVID-19, financial variables tend to capture

Findings of the Association for Computational Linguistics: EACL 2024, pages 8289–8300

backward-looking performance that may not apply in a new economic regime, while the textual commentary continues to provide forward-looking content (§6.2).

However, most work in financial prediction focuses on the most recently reported financial results and textual documents in isolation, without consideration of the temporal context of their historical patterns. While the most recent data point may be the most important, the historical context provides the ability to contextualize the current value and measure trends over time that help better predict future performance, as evidenced in both financial metrics and executive language patterns (Akbas et al., 2017; Huang et al., 2019; Cohen et al., 2020).

While financial analysts consume this data to make their quarterly earnings forecasts, it is difficult to quantitatively reconcile this complex, diverse information across sources and modalities to arrive at accurate estimates. While financial analyst estimates of company earnings are generally regarded as market expectations, they have been shown to exhibit predictable biases that can be exploited by machine-based methods (De Silva and Thesmar, 2021; Van Binsbergen et al., 2023). For instance, financial analysts often do not fully incorporate the subtle signals in macroeconomic shocks (Ball and Ghysels, 2018) or long financial documents (Frankel et al., 2018; Koval et al., 2023) into their earnings estimates. Therefore, we hypothesize that it is possible to use machine learning to effectively analyze this multimodal data and learn complex interactions between disparate sources of information that can be used to forecast earnings surprises months in advance of the report date.

In this work, we consider a multimodal time series forecasting problem in which we investigate the value of financial text, time series, and their interaction in predicting next quarter's company performance relative to market expectations. In doing so, we introduce a new multimodal time series dataset and challenging financial prediction task, and propose a novel method that learns rich temporal and cross-channel features from the numerical and textual content of noisy financial time series. In summary, we make the following contributions:

1. We design a novel multimodal time series prediction task that spans different tabular and textual financial time series from diverse sources (§4, §3, Appendix A).
2. We propose a multi-stage training process to effectively extract predictive long context embeddings from long financial documents and demonstrate that these text-based features add significant value to traditional financial variables in forecasting future company performance, particularly during periods of economic shocks, which we attribute to their ability to capture forward-looking content and contextualize historical behavior (§5.2, Table 2).
3. We systematically explore the value of each modality and temporal context across financial time series and propose a simple yet effective method that allows modality-specific temporal dynamics while still capturing complex cross-modal patterns that outperforms existing methods on this challenging task (§5).
4. We probe our proposed method through quantitative and qualitative interpretability methods to reveal insights into the value of each modality and our proposed multimodal method.
5. We demonstrate the economic value of our model predictions in a real-world trading setting with portfolio simulations that result in economically and statistically significant gains in investment performance (§6.6).

Broader Impact We hope this work will inspire future research in multimodal time series modeling with textual and tabular data from different sources, particularly as the context length and multimodal capabilities of LLMs continues to grow, as our findings suggest that small yet specialized finetuned models currently outperform LLMs on this task. We release the dataset and code at: `https://github.com/rosskoval/multimodal_ts_ff`.

2 Related Work

2.1 Text Embeddings

In the broader NLP literature, there has been great interest recently in extending the context length of Transformer-based language models (Dai et al., 2019; Beltagy et al., 2020; Zaheer et al., 2020; Guo et al., 2022; Jiang et al., 2023). However, most of these methods are not well suited for producing semantic document embeddings that either perform well for zero-shot feature extraction or provide a strong parameter initialization for downstream finetuning. While there are many pretrained models and training methods proposed for semantic text embeddings, such as SBERT (Reimers and Gurevych, 2019), SimCSE (Gao et al., 2021), and

DiffCSE (Chuang et al., 2022), most of them have a short context length intended for sentence or paragraph-level tasks. Recently, Wang et al. (2023) demonstrate strong performance with a contrastive method for finetuning long context LLMs across diverse tasks with instructions. However, they mostly evaluate their model on short text tasks with the exception of synthetic retrieval.

2.2 Financial Prediction

There has been an extensive set of company financial variables derived from their accounting statements and stock market behavior that have been found to predict future firm performance (Novy-Marx, 2013; Chordia and Shivakumar, 2006; Fama and French, 2015a; Gu et al., 2020; Chen et al., 2021). Further, it has also been found that the text of company financial documents, such as earnings calls and financial reports, contains signal that is predictive of future company performance (Kogan et al., 2009; Loughran and Mcdonald, 2011; Larcker and Zakolyukina, 2012; Cohen et al., 2020; Koval et al., 2023), but most work has focused primarily on the most recent document in isolation without context of related documents. However, Koval et al. (2024) found benefit in aligning consecutive financial reports to identify the most salient business risks. Other work has combined the textual reports with multimodal data, such as audio, graphs, videos, and tabular features to enhance predictions (Qin and Yang, 2019; Sang and Bao, 2022; Sawhney et al., 2020; Feng et al., 2021; Alanis et al., 2022; Mathur et al., 2022a; Ang and Lim, 2022; Mathur et al., 2022b). Similarly, related work has found benefit in modeling cross-channel information from related time series across different sources, such as Google Trends & Weather, for predicting key financial indicators (Zhou et al., 2020; Cao et al., 2023). However, our focus in this work is on jointly modeling the time series of paired textual financial documents and their corresponding tabular financial variables from the same firm; we do not focus on the cross-firm relationships (Ang and Lim, 2022).

3 Problem Statement

3.1 Task Formulation

We propose a multimodal financial time series task in which we predict a company's next quarter's earnings surprise (ES) from the time series of their financial, textual, and macroeconomic data. We believe these modalities and sources capture the information available to financial analysts when making their earnings forecasts, so our experiments present a unique exercise between the forecasting ability of human experts and machines. We select the Standardized Unexpected Earnings (Latane and Jones, 1979) as our measure of earnings surprise, which represent the operating performance of a company relative to market expectations and are highly followed by equity investors (Doyle et al., 2006). It is important to note that there is roughly a 3-month time horizon between prediction date and the report date, making this long horizon prediction task particularly challenging.

$$\mathrm{ES}_t = \frac{\mathrm{RepEPS_t} - \mathrm{Avg(EstEPS_t)}}{\mathrm{Std(EstEPS_t)}}$$

Since the measure is a continuous variable with approximate normal distribution, we use Mean Squared Error (MSE) as the loss function and to evaluate performance.

Multimodal Inputs We use data from a variety of modalities and sources, described in detail below, to understand the current state of a company from different perspectives. For all variables, we use quarterly data aligned with each company's reporting period and the values of the previous 8 quarters as the input time series. In this work, we refer to data source and modality interchangeably.

3.2 Autoregressive Variables (AR)

The persistence of earnings surprises is well documented in the financial literature (Kama, 2009; Loh and Warachka, 2012) with streaks found to persist up to 12 quarters, so we use the historical values of the target variable **AR** as an input.

$$\mathrm{AR_t} = [\mathrm{AR}_{t-8}, ..., \mathrm{AR}_{t-1}] \in \mathbb{R}^{1\times 8}$$

3.3 Financial Variables (FIN)

Additionally, we include the time series of 15 well-documented firm-level financial characteristics **FIN** from the asset pricing literature (Fama and French, 2015b; Cohen et al., 2020; Alanis et al., 2022; Swade et al., 2023) derived from a company's accounting statements and stock price behavior, detailed in Appendix A, including valuation, profitability, growth, volatility, and momentum. We compute these variables on a quarterly basis and normalize them to have zero mean and unit variance. While this set is not exhaustive, recent work (Swade et al., 2023) has showed that it

largely spans the principal components of the full set of firm-level financial characteristics.

$$\mathrm{FIN_t} = [\mathrm{FIN}_{t-8}, ..., \mathrm{FIN}_{t-1}] \in \mathbb{R}^{15\times 8}$$

3.4 Earnings Conference Calls (ECT)

We include the text of quarterly earnings conference calls in which company executives discuss the recent performance of the firm, their prospects, and answer questions from financial analysts covering their firms. These calls provide a rare opportunity to detect diverse signals, which can vary from clear sentiment to more subtle signs of deception (Larcker and Zakolyukina, 2012) or obfuscation (Bushee et al., 2018), that may reveal important information about the current and future prospects of the company.

$$\mathrm{ECT_t} = [\mathrm{ECT}_{t-8}, ..., \mathrm{ECT}_{t-1}]$$

3.5 Quarterly Financial Reports (QFR)

We include the text of the Management Discussion and Analysis (MDA) section from the quarterly reports (10Q/10K) of US-based public companies. This section is intended to provide management's perspective on the business results of the past year and their future prospects for the upcoming year, including information about key business risks. [1] While there are other sections, we choose to focus on the MDA because it reflects a direct communication from company management to its shareholders.

$$\mathrm{QFR_t} = [\mathrm{QFR}_{t-8}, ..., \mathrm{QFR}_{t-1}]$$

3.6 Macroeconomic State (MAC)

Finally, we include the time series of macroeconomic data (**MAC**) that represents the broader state of the US economy at each point in time to capture exogenous demand and supply-side shocks with varying impact per industry. Theses monthly indices include the following variables: Industrial Production, Inflation, Consumer Sentiment, 3M Treasury Yield, 10YR Treasury Yield, Term Spread, Default Spread, and Oil Prices. Following Ball and Ghysels (2018), we compute quarterly growth rates for each variable to arrive at our macroeconomic time series of 8 channels.

$$\mathrm{MAC_t} = [\mathrm{MAC}_{t-8}, ..., \mathrm{MAC}_{t-1}] \in \mathbb{R}^{6\times 8}$$

We merge this time series to align with each company's prior fiscal quarter end date.

[1] https://www.sec.gov/files/reada10k.pdf

4 Data

	Train	Validation	Test
Start Date	Jan-2010	Jan-2016	Jan-2017
End Date	Dec-2015	Dec-2016	Dec-2020
# Samples	7,188	1,362	5,723
# Firms	710	638	708
# Modalities	4	4	4
# Time Steps	8	8	8

Table 1: Summary Statistics on each sample split.

Data Acquisition and Curation We source Reported Earnings per Share (EPS) and Analyst Consensus Estimates of EPS from FactSet Fundamentals and Consensus Estimates, respectively, to compute the Earnings Surprise target variable. We collect English conference calls from FactSet Document Distributor. We source quarterly and annual reports from Notre Dame Software Repository for Accounting and Finance. We collect the monthly macroeconomic indices from the ST. Louis FED. Please see §A.2 for further details on the extensive data curation process.

Data Statistics and Task Formulation We focus our analysis on the largest publicly trade companies in the US (MSCI USA Index) and require that each sample point possess valid data for each input variable across historical time steps. Then, we temporally partition the data into train (2010-2015), validation (2016), and test (2017-2020) sets. We provide summary statistics in Table 1.

5 Methods

We systematically explore a comprehensive set of unimodal baselines and propose a novel multimodal method that significantly outperforms the best unimodal models.

5.1 Multivariate Time Series

Since each of our modalities constitutes a multivariate time series with complex cross-channel interactions, we consider Time Series Mixer (Chen et al., 2023) as our backbone time series encoder because of its strong performance on multichannel time series forecasting. TSM is a simple yet effective all-MLP neural architecture that performs alternating forms of feature mixing across the time (TM) and channel dimension (CM) in each layer of the network. We denote the time series model as **TSM**, with input $X \in \mathbb{R}^{T\times C}$, T time steps,

C channels, and returns features $H \in \mathbb{R}^{T \times C}$ after interaction across the temporal and channel dimensions, expressed below:

$$\mathrm{TM}_l(X) = \mathrm{BN}(X + \sigma(W_{\mathrm{TM}} X_{*,t} + b_{\mathrm{TM}}))$$
$$\mathrm{CM}_l(X) = \mathrm{BN}(X + \sigma(W_{\mathrm{CM}} X_{c,*} + b_{\mathrm{CM}}))$$
$$\mathrm{TSM}_l(X) = \mathrm{CM}_l(\mathrm{TM}_l(X))$$

We perform ablations of this choice in §6.5. In each modality time series, we use the last $T = 8$ quarters of values as inputs. We conduct a grid search over the model architecture hyperparameters detailed in §A.6.

5.2 Multi-Stage Textual Feature Extraction (MTFE)

Since ECTs and QFRs are long textual documents (5K+ tokens each) and thus cannot be concatenated over the time series (50K+ tokens) and trained end-to-end with limited computational resources, we propose a novel multi-stage training method to extract predictive features from them. It is important to note that most of the text embedding literature (Muennighoff et al., 2022) has focused on short-context texts at the sentence or paragraph-level, such as consumer reviews, which contains distinct characteristics from our problem setting.

We initialize our text encoder with the pretrained BigBird (Zaheer et al., 2020) checkpoint due to its long context length and strong performance in long document understanding tasks. This process consists of the following steps.

(1) **Multitask Domain Adaptation**: Domain adaptation is important to the success of using pretrained language models for domain-specific text (Han and Eisenstein, 2019; Gururangan et al., 2020). Since we believe our tasks require a specialized understanding of financial language, we conduct multi-task domain-adaptive pretraining (**MTDA**).

To do so, we adapt the BigBird-base model (Zaheer et al., 2020) to the financial domain by jointly performing long context masked language modeling (**LC-MLM**) and long-context DiffCSE (**LC-DiffCSE**), in which we adapt the contrastive pretraining framework DiffCSE (Chuang et al., 2022) designed for short-context models to long-context BigBird. We do this by prepending and assigning global attention to the original document embedding in the replaced token detection objective to encourage the model to use that information to predict the replaced tokens, resulting in more fine-grained document representations. We conduct this pretraining of over a pooled corpus of in-domain ECTs and QFRs that occur during the training date period for a total of 25K training steps.

This multi-task pretraining process serves several purposes. Firstly, **LC-MLM** adapts the model to the complex language of the financial domain. Secondly, **LC-DiffCSE** learns to produce aggregated document representations that are sensitive to topically similar but semantically different financial text. We believe both of these steps are critical towards capturing strong document representations and we demonstrate their value in Table 3.

(2) **Supervised Finetuning (SFT)**: We finetune the adapted model on the text of the last time step ECT and QFR (single document) and the corresponding earnings surprise measure to learn task-specific features.

$$\hat{Y}_t = \mathrm{ENC}_{\mathrm{ECT}}(\mathrm{ECT}_t), \mathrm{ENC}_{\mathrm{QFR}}(\mathrm{QFR}_t)$$

(3) **Feature Extraction**: We extract features from the last layer embeddings E of the [CLS] token from the finetuned encoder ENC for each document.

$$E_{\mathrm{ECT}} = \mathrm{ENC}_{\mathrm{ECT}}(\mathrm{ECT}); E_{\mathrm{QFR}} = \mathrm{ENC}_{\mathrm{QFR}}(\mathrm{QFR})$$

These features contain richer information than solely the output predictions, allowing the time series model to capture multifaceted trends in executive language patterns over time. We compare our approach with existing pretrained embedding models in §6.3.

5.3 Multimodal Fusion Methods

We explore three multimodal fusion methods to mix the time series across modalities with varying channel dimensionality, using TSM as the time series encoder.

There are two paradigms of approaches in multivariate time series forecasting and we explore them both here as baselines. Firstly, Channel-Mixing (Chen et al., 2023), in which all univariate channels are concatenated together and treated as a single multi-channel signal, assumes that there exists cross-channel information. In this case, since each input modality contains a different number of dimensions, we project them all into the same vector dimension and space before concatenating. We label this approach cross-modality mixer **CMM**.

Alternatively, Channel-Independence assumes that the relationship between each univariate time series is independent and should be processed separately but with shared model weights across channels (Nie et al., 2022). This approach has found success particularly when the cross-channel signal is weak, and thus cross-channel interaction leads to overfitting. We apply this approach with the original model (**PatchTST**) and an architecture-controlled version (TSMixer), labeled modality independent, channel independent (**MICI**).

However, we believe that these two baseline approaches are not well suited to this multimodal problem for two reasons. Firstly, we hypothesize that each modality contains rich cross-channel patterns that Channel-Independence cannot capture. Secondly, we believe that differences in the temporal dynamics of each modality could lead to statistical noise and incompatibilities if relying on a single model, and that mixing modalities too early in the feature learning process is likely to lead to overfitting.

Therefore, we propose an different approach that imposes this inductive bias and treats each modality (**AR+FIN**, **ECT**, **QFR**, **MAC**) as a separate multi-channel time series. In our modality-independent, channel mixing method (**MICM**), we first process each modality multi-channel time series independently using modality-adaptive time series models and then linearly project them into the same space and dimension.

$$X_{\text{AR+FIN}} = W_{\text{AR+FIN}}\text{TSM}_{\text{AR+FIN}}([\text{AR};\text{FIN}])$$

$$\begin{aligned} X_{\text{ECT}} &= W_{\text{ECT}}\text{TSM}_{\text{ECT}}(\text{E}_{\text{ECT}}) \\ X_{\text{QFR}} &= W_{\text{QFR}}\text{TSM}_{\text{QFR}}(\text{E}_{\text{QFR}}) \\ X_{\text{MAC}} &= W_{\text{MAC}}\text{TSM}_{\text{MAC}}(\text{MAC}) \end{aligned}$$

This approach allows the the modality-specific models to learn cross-time and cross-channel patterns that are adaptive to each modality and produce features that can be fused across modalities at a later stage. We concatenate the resulting mixed features together and introduce a cross-modality fusion (**CMF**) module to interact cross-modality channels based upon temporal similarity. This module consists of a cross-channel multihead attention mechanism (MHA) and layer normalization (LN) with residual connections.

$$X_{MM} = [X_{\text{AR+FIN}}; X_{\text{ECT}}; X_{\text{QFR}}; X_{\text{MAC}}]$$

$$X_O = \text{LN}(X_{MM} + \text{MHA}_{\text{c},:}(X_{MM}, X_{MM}, X_{MM}))$$

Again, this design imposes the inductive bias that while cross-modality channel patterns exist, they should be carefully interacted late in the network base upon temporal similarity in a learned projection space. We flatten the output and include a 2-layer MLP head for prediction:

$$\hat{Y}_t = \text{MLP}(\text{Flatten}(X_O))$$

We conduct a grid search over the TSM model architecture over each modality in isolation (unimodal), detailed in Appendix A, allowing each modality-specific time series model to have a different representational capacity to adapt to varying complexities of temporal and cross-channel patterns.

6 Experimental Results and Analysis

6.1 Value of Multiple Modalities

The results in Table 2 highlight the challenging nature of the task, but **we find broad consistency in the relative performance of each method across modalities and time periods**.

We find that **all input features benefit from including the time series context in addition to the most recent time step** albeit modestly. While the benefit is modest in magnitude, we note the consistency in positive improvements across modalities and time periods. We find that the QFRs tend to benefit the most from this temporal context. This is result is consistent with recent work (Cohen et al., 2020) because there is considerable boilerplate content in financial reports that does not vary much year to year, making it difficult to identify new information, and therefore requiring the historical context to identify and contextualize the differences over time.

We also find that **both sources of textual data provide considerable value beyond the financial time series variables**. However, we find that the marginal benefit of ECTs is greater than that of the QFRs. We suspect this is because the ECTs contain richer information in the form of both less scripted language by the company executives and the inclusion of analyst question-answer exchanges that provide an additional perspective of analyst context.

We conjecture that the complementary nature of the textual and tabular data is partly due to the fact that tabular data captures the past performance while the text-based models contextualize the persistence of that performance with qualitative in-

Input	Modality	Time Steps	MSE_{2017}	MSE_{2018}	MSE_{2019}	MSE_{2020}	MSE
AR	TABULAR	1	1.33	1.52	1.60	2.50	1.77
AR	TABULAR	8	1.30	1.50	1.58	2.45	1.71
FIN	TABULAR	1	1.33	1.51	1.60	2.47	1.77
FIN	TABULAR	8	1.29	1.48	1.56	2.44	1.72
ECT	TEXT	1	1.19	1.46	1.47	2.35	1.62
ECT	TEXT	8	1.18	1.46	1.47	2.29	1.60
QFR	TEXT	1	1.27	1.46	1.64	2.29	1.68
QFR	TEXT	8	1.27	1.44	1.52	2.18	1.62
AR + FIN	TABULAR	8	1.22	1.38	1.42	2.29	1.60
AR + FIN + ECT*	MULTIMODAL	8	1.17	1.34	1.35	2.04	1.48*
AR + FIN + ECT + QFR**	MULTIMODAL	8	1.14	1.30	1.31	1.96	1.44**
AR + FIN + ECT + QFR + MAC	MULTIMODAL	8	**1.13**	**1.28**	**1.30**	**1.92**	**1.42**

Table 2: Main Results (smaller is better, **best in bold**): Model performance on the test set of our multimodal Earnings Surprise Prediction task. All results use our proposed multimodal method, including modality-specific **TSMixer** encoders, **MTFE** to extract textual features, and cross-modality fusion **CMF** to mix modalities. "Time Steps" indicate how many quarters of data are used in the time series model. **AR** = autoregressive earnings surprise lags, **FIN** = tabular financial variables, **ECT** = text features from earnings call transcripts, **QFR** = text features from quarterly financial reports, and **MAC** = Macroeconomic indicator variables. *, ** indicates the performance of the specified model is statistically better ($p < 0.05$) than that of the next best performing model on the test set according to the Wilcoxon Signed-Rank Test.

formation and augment it with forward-looking content, which we investigate further in the next section.

6.2 Case Study and Qualitative Analysis

To further understand the value of the textual modalities, we analyze model behavior during the 2020 time period, which was characterized by a significant economic shock caused by the COVID-19 pandemic. While we observe larger errors during the period, we also observe the largest improvement from the incorporation of text-based time series. We believe the value of Text to be most significant in the 2020 (COVID-19) regime shift likely because the executive commentary contained in the text contains forward-looking content about the future prospects of the company in a new economic regime while **AR+FIN** largely captures historical behavior under an old economic regime. For example, it is well known that Internet Technology companies performed well during this period as people were confined to their homes and more active on the internet, while Hotels & Restaurants struggled. This regime shift would not be captured in the historical time series of company financial performance but would be reflected in the executive discussions contained in their forward-looking disclosures. We confirm this by comparing the differences between model predictions for Q2-2020. We find that the models with text-based inputs (**AR+FIN+ECT+QFR**) have 0.47 higher average prediction values for companies in the Internet Technology industry than those that only rely on financial variables (**AR+FIN**). Conversely, we find that the models with text-based inputs have 0.38 lower average prediction values for companies in the Hotels & Restaurants industry than those that only rely on financial variables.

6.3 Text Encoding Methods

In Table 3, we compare our **MTFE** approach with strong pretrained baselines to demonstrate the value of our approach. Firstly, we compare with **Mistral-Embedding**, which is a state-of-the-art finetuned embedding checkpoint (Wang et al., 2023) of the Mistral-7B model (Jiang et al., 2023), a decoder-only language model that supports long contexts. The model has been finetuned with contrastive learning on a collection of document pairs across diverse tasks. We also include a version **E5-LongEmbed** of the weakly supervised embedding model E5 (Wang et al., 2022) that was optimized for long context lengths (Zhu et al., 2024). Secondly, we also compare with short-context models, including **SBERT** (Reimers and Gurevych, 2019), which was contrastively pretrained on a massive corpus of 1B weakly supervised text pairs, as well as with domain-specific language model **FinBERT** (Huang et al., 2022). We apply the short context models at the sentence level and average the resulting embeddings over all sentences in each document. Finally, we include a pretrained financial

sentiment classifier **FinBERT-Sent** (Araci, 2019) applied at the sentence-level (Alanis et al., 2022), detailed in Appendix A.

We find that BigBird with **MTFE** provides considerable improvement over these SOTA pretrained models, suggesting that the task requires a specialized understanding of financial language, task-specific predictive features, and document-level context. This improvement over **Mistral-Embedding** suggests that smaller encoder-only models with specialized training can outperform much larger, general-purpose decoder-only LLMs on domain-specific tasks, consistent with the recent findings of Shah and Chava (2023). We conjecture that the benefits of **MTFE** arises from a combination of specialized domain and task adaptation to capture subtle signals, and a bidirectional encoding that provides better relative importance estimation within a long context (Liu et al., 2023).

Text Encoder	Params	ECT	QFR
MTFE	150M	**1.60***	**1.62***
w/o LC-DiffCSE	150M	1.64	1.65
w/o LC-MLM	150M	1.66	1.68
w/o SFT	150M	1.75	1.77
FinBERT-Sent	130M	1.91	1.94
SBERT	120M	1.72	1.75
FinBERT	130M	1.75	1.77
Mistral-Embedding	7B	1.70	1.72
E5-LongEmbed	110M	1.70	1.73

Table 3: Results indicate MSE on the test set using different textual encoding methods for feature extraction from the specified financial documents and demonstrate the value of each step of our MTFE process. * indicates the performance of the specified model is statistically better ($p < 0.05$) than that of the next best performing model on the test set according to the Wilcoxon Signed-Rank Test.

6.4 Multimodal Fusion Methods

In Table 4, we our compare modality-independent, channel mixing with late stage cross-modality fusion method **MICM** with the cross-modality channel concatenation **CMM** (typically used in multivariate time series forecasting), and channel independence across modalities **MICI**. While all multimodal methods outperform the unimodal baselines, we find that our proposed method **MICM** performs significantly better than **CMM** and **MICI**. We believe this result is due to: (1) **differential temporal patterns across modalities and the resulting incongruence that results from treating them equally (CMM)**; (2) **the rich cross-channel interactions that exists within each modality time series that independence ignores (MICI)**.

6.5 Time Series Encoder

In Table 4, we ablate our choice of TSMixer as our time series encoder with generic neural models (MLP, Transformer) and channel-independent, time series model PatchTST to further justify our use of a specialized temporal model with cross-channel interaction in the proposed method. We find that TSMixer demonstrates strong performance on this task due to its ability to both capture intra-channel temporal patterns as well as cross-channel interactions.

Method	MSE
TSMixer-MICM	**1.42***
TSMixer-CMM	1.49
TSMixer-MICI	1.47
MLP	1.57
Transformer	1.50
PatchTST	1.48

Table 4: Comparison of the performance of using different time series encoders and modality fusion methods. * indicates the performance of the specified model is statistically better ($p < 0.05$) than that of the next best performing model according to the Wilcoxon Signed-Rank Test.

6.6 Portfolio Simulations

In Table 5, we demonstrate the economic value of our model predictions using portfolio simulations. We form monthly long-short (*market-neutral*) quintile portfolios (Fama and French, 2015a) by sorting stocks based on (test set) model predictions, detailed in §A.1. We follow Cong et al. (2021) in reporting net portfolio performance that includes conservative estimates of the impact of transaction costs on portfolio implementation. The resulting performance of our proposed method generates strong investment performance that is that is economically and statistically better than the tabular financial time series baseline. These results demonstrate that the model predictions contain significant signal that is not priced into the market with substantial value in a real-world trading setting.

7 Conclusion

In conclusion, we introduce a new multimodal financial time series prediction task. We propose a novel textual feature extraction process and mul-

Statistic	AR	+FIN+MAC	+ECT+QFR
Net Return	4.51	7.08	9.74
Volatility	9.64	9.68	9.47
Net Sharpe Ratio	0.47	0.73	1.03

Table 5: 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.1.

timodal fusion method that demonstrates state-of-the-art performance on this challenging task. Our extensive experimental results and interpretability analysis reveal insights into the value of each modality and our proposed method. Notably, we find that the greatest gains in performance are achieved when jointly considering both the temporal and cross-modal context that are not possible with either context alone.

Limitations

While we demonstrate that small encoder-only language models with in-domain, task specialized training can outperform SOTA LLMs in producing long text embeddings for this task, we acknowledge that there may exist better prompts or other ways to use LLMs in the textual feature extraction process and leave it to future work to perform more extensive prompt engineering on the task instructions.

Our experiments also demonstrate that the inclusion of text and time series data improves the ability of the model to predict future earnings surprises, and that those improvements translate to gains in simulated investment performance. However, we also acknowledge that the simulated investment performance does not include the impact of transaction costs from portfolio turnover, which can erode gains if not carefully managed, and therefore does not necessarily directly translate to a live trading setting. We leave it to future work to assess their utility in real-world portfolio management.

Ethics Statement

We acknowledge that the multimodal financial time series dataset used in this work only contains English text from the largest US-based companies so it is possible that some populations may be underrepresented in this sample. We hope to be able to extend this work to international companies and financial text written in other languages in the future.

Acknowledgements

We would like to thank AJO Vista and FactSet for providing access to and permission to release the data. The authors are solely responsible for the content and views expressed in this publication and do not reflect those of the affiliated institutions.

References

Ferhat Akbas, Chao Jiang, and Paul D Koch. 2017. The trend in firm profitability and the cross-section of stock returns. *The Accounting Review*, 92(5):1–32.

Emmanuel Alanis, Sudheer Chava, and Agam Shah. 2022. Benchmarking machine learning models to predict corporate bankruptcy. *arXiv preprint arXiv:2212.12051*.

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.

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

Ryan T Ball and Eric Ghysels. 2018. Automated earnings forecasts: Beat analysts or combine and conquer? *Management Science*, 64(10):4936–4952.

Iz Beltagy, Matthew E Peters, and Arman Cohan. 2020. Longformer: The long-document transformer. *arXiv preprint arXiv:2004.05150*.

Stephen Brown, Stephen A Hillegeist, and Kin Lo. 2004. Conference calls and information asymmetry. *Journal of Accounting and Economics*, 37(3):343–366.

Brian J Bushee, Ian D Gow, and Daniel J Taylor. 2018. Linguistic complexity in firm disclosures: Obfuscation or information? *Journal of Accounting Research*, 56(1):85–121.

Defu Cao, Yixiang Zheng, Parisa Hassanzadeh, Simran Lamba, Xiaomo Liu, and Yan Liu. 2023. Large scale financial time series forecasting with multi-faceted model. In *Proceedings of the Fourth ACM International Conference on AI in Finance*, pages 472–480.

Liping Chen, Yan Deng, Xi Wang, Frank K. Soong, and Lei He. 2021. Speech bert embedding for improving prosody in neural tts. volume 2021-June.

Si-An Chen, Chun-Liang Li, Nate Yoder, Sercan O Arik, and Tomas Pfister. 2023. Tsmixer: An all-mlp architecture for time series forecasting. *arXiv preprint arXiv:2303.06053*.

Tarun Chordia and Lakshmanan Shivakumar. 2006. Earnings and price momentum. *Journal of financial economics*, 80(3):627–656.

Yung-Sung Chuang, Rumen Dangovski, Hongyin Luo, Yang Zhang, Shiyu Chang, Marin Soljačić, Shang-Wen Li, Scott Yih, Yoon Kim, and James Glass. 2022. Diffcse: Difference-based contrastive learning for sentence embeddings. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4207–4218.

Lauren Cohen, Christopher Malloy, and Quoc Nguyen. 2020. Lazy prices. *The Journal of Finance*, 75(3):1371–1415.

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*.

Zihang Dai, Zhilin Yang, Yiming Yang, Jaime G Carbonell, Quoc Le, and Ruslan Salakhutdinov. 2019. Transformer-xl: Attentive language models beyond a fixed-length context. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2978–2988.

Tim De Silva and David Thesmar. 2021. Noise in expectations: Evidence from analyst forecasts. Technical report, National Bureau of Economic Research.

Jeffrey T Doyle, Russell J Lundholm, and Mark T Soliman. 2006. The extreme future stock returns following i/b/e/s earnings surprises. *Journal of Accounting Research*, 44(5):849–887.

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

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

Qi Feng, Han Chen, and Ruohan Jiang. 2021. Analysis of early warning of corporate financial risk via deep learning artificial neural network. *Microprocessors and Microsystems*, 87.

Richard M. Frankel, Jared N. Jennings, and Joshua A. Lee. 2018. Using natural language processing to assess text usefulness to readers: The case of conference calls and earnings prediction. *SSRN Electronic Journal*.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. 2021. Simcse: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910.

Shihao Gu, Bryan Kelly, and Dacheng Xiu. 2020. Empirical asset pricing via machine learning. *The Review of Financial Studies*, 33(5):2223–2273.

Mandy Guo, Joshua Ainslie, David C Uthus, Santiago Ontanon, Jianmo Ni, Yun-Hsuan Sung, and Yinfei Yang. 2022. Longt5: Efficient text-to-text transformer for long sequences. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 724–736.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. 2020. Don't stop pretraining: Adapt language models to domains and tasks. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8342–8360.

Xiaochuang Han and Jacob Eisenstein. 2019. Unsupervised domain adaptation of contextualized embeddings for sequence labeling. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 4238–4248.

Allen H Huang, Hui Wang, and Yi Yang. 2022. Finbert: A large language model for extracting information from financial text. *Contemporary Accounting Research*.

Dashan Huang, Huacheng Zhang, and Guofu Zhou. 2019. Twin momentum: Fundamental trends matter. *Available at SSRN 2894068*.

Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. 2023. Mistral 7b. *arXiv preprint arXiv:2310.06825*.

Itay Kama. 2009. On the market reaction to revenue and earnings surprises. *Journal of Business Finance & Accounting*, 36(1-2):31–50.

Shimon Kogan, Dimitry Levin, Bryan R. Routledge, Jacob S. Sagi, and Noah A. Smith. 2009. Predicting risk from financial reports with regression.

Ross Koval, Nicholas Andrews, and Xifeng Yan. 2023. Forecasting earnings surprises from conference call transcripts. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8197–8209, Toronto, Canada. Association for Computational Linguistics.

Ross Koval, Nicholas Andrews, and Xifeng Yan. 2024. Learning to compare financial reports for financial forecasting. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 500–512, St. Julian's, Malta. Association for Computational Linguistics.

David F Larcker and Anastasia A Zakolyukina. 2012. Detecting deceptive discussions in conference calls. *Journal of Accounting Research*, 50(2):495–540.

Henry A Latane and Charles P Jones. 1979. Standardized unexpected earnings–1971-77. *The journal of Finance*, 34(3):717–724.

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

Roger K Loh and Mitch Warachka. 2012. Streaks in earnings surprises and the cross-section of stock returns. *Management Science*, 58(7):1305–1321.

Tim Loughran and Bill Mcdonald. 2011. When is a liability not a liability? textual analysis, dictionaries, and 10-ks. *Journal of Finance*, 66.

Puneet Mathur, Mihir Goyal, Ramit Sawhney, Ritik Mathur, Jochen L Leidner, Franck Dernoncourt, and Dinesh Manocha. 2022a. Docfin: Multimodal financial prediction and bias mitigation using semi-structured documents. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1933–1940.

Puneet Mathur, Atula Neerkaje, Malika Chhibber, Ramit Sawhney, Fuming Guo, Franck Dernoncourt, Sanghamitra Dutta, and Dinesh Manocha. 2022b. Monopoly: Financial prediction from monetary policy conference videos using multimodal cues. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 2276–2285.

Niklas Muennighoff, Nouamane Tazi, Loïc Magne, and Nils Reimers. 2022. Mteb: Massive text embedding benchmark. *arXiv preprint arXiv:2210.07316*.

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*.

Robert Novy-Marx. 2013. The other side of value: The gross profitability premium. *Journal of financial economics*, 108(1):1–28.

Yu Qin and Yi Yang. 2019. What you say and how you say it matters: Predicting stock volatility using verbal and vocal cues. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 390–401.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Yunxin Sang and Yang Bao. 2022. DialogueGAT: A graph attention network for financial risk prediction by modeling the dialogues in earnings conference calls. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 1623–1633, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Ramit Sawhney, Piyush Khanna, Arshiya Aggarwal, Taru Jain, Puneet Mathur, and Rajiv Shah. 2020. Voltage: Volatility forecasting via text audio fusion with graph convolution networks for earnings calls. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)*, pages 8001–8013.

Agam Shah and Sudheer Chava. 2023. Zero is not hero yet: Benchmarking zero-shot performance of llms for financial tasks. *arXiv preprint arXiv:2305.16633*.

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

Jules H Van Binsbergen, Xiao Han, and Alejandro Lopez-Lira. 2023. Man versus machine learning: The term structure of earnings expectations and conditional biases. *The Review of Financial Studies*, 36(6):2361–2396.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022. Text embeddings by weakly-supervised contrastive pre-training. *arXiv preprint arXiv:2212.03533*.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. 2023. Improving text embeddings with large language models. *arXiv preprint arXiv:2401.00368*.

Manzil Zaheer, Guru Guruganesh, Avinava Dubey, Joshua Ainslie, Chris Alberti, Santiago Ontanon, Philip Pham, Anirudh Ravula, Qifan Wang, Li Yang, and Amr Ahmed. 2020. Big bird: Transformers for longer sequences. volume 2020-December.

Dawei Zhou, Lecheng Zheng, Yada Zhu, Jianbo Li, and Jingrui He. 2020. Domain adaptive multi-modality neural attention network for financial forecasting. In *Proceedings of The Web Conference 2020*, pages 2230–2240.

Dawei Zhu, Liang Wang, Nan Yang, Yifan Song, Wenhao Wu, Furu Wei, and Sujian Li. 2024. Longembed: Extending embedding models for long context retrieval. *arXiv preprint arXiv:2404.12096*.

A Appendix

A.1 Portfolio Simulations

In Table 5, we demonstrate the economic value of our model predictions using portfolio simulations. We form monthly long-short (market-neutral) quintile portfolios (Fama and French, 2015a) by sorting stocks based on (test set) model predictions from the most recently reported quarter, and buying those in the top 20% and shorting those in the bottom 20% on a monthly basis in equal proportions.

In Table 5, we report net portfolio performance that includes 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, to compute net returns. Since the data is updated quarterly and the investment universe consists of large, liquid US stocks, the proposed trading strategy is likely to have relatively low turnover and incur very modest transaction costs.

A.2 Data Curation

We merge all data across modalities to align with the quarterly fiscal period end dates for each company. Therefore, for each Earnings Surprise forecast quarter date (e.g. Q2 2020), we select data that was available within 5 business days of the end of the previous quarter fiscal end date, including all financial variables, quarterly financial reports, earnings conference calls, and macroeconomic data. Although some of this data is available at a higher frequency than once a quarter, such as stock price-based financial variables or monthly economic indicators, we only select the values of such data that were available as of 5 business days after the last fiscal period end date. Thus, our predictions are always made 3 months in advance of the earnings report date. We leave it to future work to explore the value in updating the predictions at a higher frequency.

A.3 Pretrained Language Models

We develop all Transformer-based models in PyTorch and source all pretrained checkpoints from HuggingFace.

A.4 Textual Feature Extraction

We include pretrained financial sentiment classifier **FinBERT-Sent + Linear** (Araci, 2019) applied at the sentence-level (Alanis et al., 2022):

$$\text{FinBERT-Sent} = \frac{\text{\#PositiveSentences - \#NegativeSentences}}{\text{\#TotalSentences}}$$

A.5 Firm Financial Variables

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). This set of variables is not exhaustive but has been shown to span the set of principle components of firm characteristics.

A.6 Training Details and Hyperparamter Tuning

We perform all experiments on a single Tesla A100 GPU with 40GB in memory. We use AdamW to optimize all parameters. We conduct an extensive grid search over the neural architecture of the time series encoder for each modality in isolation, including number of layers $\{1, 2, 4, 6, 8, 10\}$ and hidden sizes $\{16, 32, 64, 128, 256, 512\}$, as well as learning rates $\{1e-5, 5e-4, 1e-4, 5e-3, 1e-3, 5e-3\}$ and batch sizes $\{32, 64, 128, 256\}$. We train all models for 20 epochs and select the best checkpoint based off validation set performance for test evaluation. For computational constraints, we train all models using mixed precision training, and apply gradient checkpointing to satisfy GPU memory constraints, and clip gradient norms.

A.7 MT-DA Pretraining Details

We conduct the MT-DA pretraining process for the document-level, BigBird backbone models for a maximum of 25K training steps or until the loss on the validation set increases, using the same hyperparameter configuration and settings as Chuang et al. (2022). This pretraining process takes multiple days of run time for each framework and indicates the difficulty of pretraining these Efficient Transformers models on domain relevant text. We adapt the DiffCSE objective to the long context BigBird model by prepending and assigning global attention to the original document embedding in the RTD objective to encourage the model to use that information to predict the replaced tokens, resulting in more fine-grained document representations. We conduct this pretraining process over a pooled corpus of in-domain ECTs and QFRs that occur during the training date period for a total of 25K training steps. We use pretrained checkpoint of BigBird-base as the fixed generator (masked language model) model because there are no widely accepted distilled or smaller versions. We tune the tradeoff between the LC-MLM and LC-DiffCSE loss weight in the multitask DA objective over $\{0.1, 0.25, 0.5, 0.75, 0.90\}$ according to validation set performance. Please see Chuang et al. (2022) for more details on the DiffCSE framework.