MONOPOLY: Financial Prediction from MONetary POLicY Conference Videos Using Multimodal Cues

Puneet Mathur
puneetm@umd.edu
University of Maryland, College Park

Atula Neerkaje
atula.neerkaje@learner.manipal.edu
Manipal Institute of Technology

Malika Chhibber
f20180611@goa.bits-pilani.ac.in
BITS Pilani, Goa Campus

Ramit Sawhney
rsawhney31@gatech.edu
Georgia Institute of Technology

Fuming Guo
fuming.guo@fmr.com
Fidelity Investments

Franck Dernoncourt
dernonco@adobe.com
Adobe Research

Sanghamitra Dutta
sanghamitra2612@gmail.com
University of Maryland, College Park

Dinesh Manocha
dmanocha@umd.edu
University of Maryland, College Park

ABSTRACT

Risk prediction and price movement classification are essential tasks in financial markets. Monetary policy calls (MPC) provide important insights into the actions taken by a country's central bank on economic goals related to inflation, employment, prices, and interest rates. Analyzing visual, vocal, and textual cues from MPC calls can help analysts and policymakers evaluate the economic risks and make sound investment decisions. To aid the analysis of MPC calls, we curate the Monopoly dataset, a collection of public conference call videos along with their corresponding audio recordings and text transcripts released by six international banks between 2009 and 2022. Our dataset is the first attempt to explore the benefits of visual cues in addition to audio and textual signals for financial prediction tasks. We introduce MPCNet, a competitive baseline architecture that takes advantage of the cross-modal transformer blocks and modality-specific attention fusion to forecast the financial risk and price movement associated with the MPC calls. Empirical results prove that the task is challenging, with the proposed architecture performing 5-18% better than strong Transformer-based baselines. We release the MPC dataset and benchmark models to motivate future research in this new challenging domain.

CCS CONCEPTS

• **Computing methodologies → Visual inspection**.

KEYWORDS

finance, monetary policy calls, multimodal learning, video analysis

ACM Reference Format:
Puneet Mathur, Atula Neerkaje, Malika Chhibber, Ramit Sawhney, Fuming Guo, Franck Dernoncourt, Sanghamitra Dutta, and Dinesh Manocha. 2022. MONOPOLY: Financial Prediction from MONetary POLicY Conference Videos Using Multimodal Cues. In *Proceedings of the 30th ACM International Conference on Multimedia (MM '22), October 10–14, 2022, Lisboa, Portugal.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3503161.3548380

1 INTRODUCTION

Predicting how the prices of a financial asset will vary over a certain period is an important financial analysis task for investors and policymakers [29]. Understanding the sentiment of the economy and it's associated risk perceptions can help analysts make better decisions about investment returns, while policymakers can implement cautionary monetary measures in order to maintain a healthy economy [10, 51]. With unparalleled advances in multimodal learning, a massive amount of unstructured data is accessible to investors for financial forecasting [25]. One such rich source of information is the Monetary Policy Conference (MPC's) call. These hour-long, public video conferences are held periodically where the governors of a country's central bank[1] (eg., the Federal Reserve Bank in the United States) meet to discuss the actions undertaken to improve the financial conditions of the country, explain their stance on the monetary policy, and assess the risks to economic growth. The MPC calls are a combination of a prepared press speech by the governor followed by a spontaneous question-answering session with the journalists [33]. The public presentation sheds light on the announcements regarding policy decisions and gives indications about the future path of the economy. The question-answer session involves the call participants like media reporters and market analysts engaging in a dialogue with the governors to analyze a range of economic factors such as inflation, employment, value of currency, stock market growth and interest rates on loans.

Prior works [9, 43] have highlighted the impact of MPC calls on financial stock markets as evident from a "higher than normal" trading across different financial assets. For instance, [20] gives an example of how the volatility of S&P 500 index can be observed to be roughly three times larger on days when the Federal Reserve Bank conducts its MPC calls in the US compared to other times.

MM '22, October 10–14, 2022, Lisboa, Portugal
© 2022 Association for Computing Machinery.
ACM ISBN 978-1-4503-9203-7/22/10…$15.00
https://doi.org/10.1145/3503161.3548380

**Opinions expressed here are the author's own, and do not represent the views of Fidelity Investments. A standard disclaimer applies.

[1] https://www.investopedia.com/terms/c/centralbank.asp

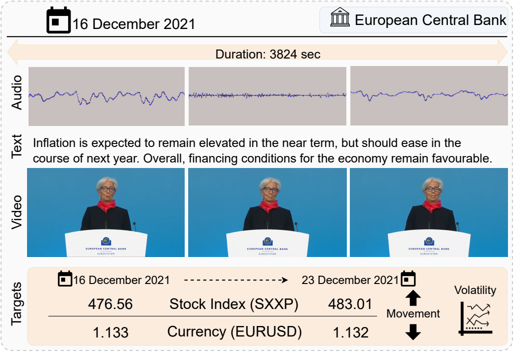


Figure 1: A sample from a Monetary Policy Call held by the Europen Central Bank. The Governor first presents a prepared press speech, followed by a spontaneous question and answer (Q&A) session with journalists. The meeting ended with an adverse market reaction that led to declining currency value and a high volatility in stock prices.

Hence, shareholders critically analyze the multimodal MPC calls to forecast stock market indices, treasury bonds, prices of gold, and currency exchange rates post the conference call [53]. Prior findings [21] suggest that the minutes of the MPC calls can provide important market-relevant information for several financial assets as mentioned in Table 1 and need to be assessed systematically.

There is anecdotal evidence that non-verbal cues such as complexity of language, vocal tone and facial expressions of the speakers can be indicative and correlated with trading activities in the financial markets [11, 30]. Although existing research has used text and audio for financial predictions [41, 42, 47, 48], use of visual cues as part of multimodal input has been largely limited. Existing NLP literature has focused on what is being said during the press conferences while there is a need to focus on how it is being said. This can accomplished by exploiting the visual aspects of the conferences for scrutinizing the human behavior such as eye-movements, facial expressions, postures, and gaits [33]. According to [57], behavioral clues may reflect emotions that subjects might want to hide. Variability across different speakers makes it extremely difficult to detect these expressions in real time. For instance, Figure 1 depicts an MPC call held by European Central bank where the tone of the conference takes a more negative turn when the governor tries to evade questions on future inflation. The textual content indicates an optimistic outlook on long-term inflation despite an overall pessimistic vocal tone. The followup discussion depicts the speaker hesitancy in indulging more details to the reporters, accompanied with facial expressions that could indicate stress. Consequently, the meeting ended with an adverse market reaction that led to declining currency value and a high volatility in stock prices. Motivated by prior works, we explore multimodal deep learning approaches that can extract complementary information from multiple modalities to improve financial modeling. Our work takes the first step in multimodal financial modeling on MPC calls by utilizing the visual, vocal, and verbal modalities simultaneously.

Our **contributions** in this work can be summarized as:

- We curate a public dataset[2], Monopoly: Monetary Policy Call Dataset, consisting of 340 video conference calls spanning over 350 hours between 2009 to 2022 extracted from 6 major English-speaking economies - USA, Canada, European Union, United Kingdom, New Zealand, and South Africa.
- We accompany the dataset with several strong neural baselines. Our proposed methodology, MPCNet utilizes video frames, audio recordings, and utterance-aligned transcripts, learnt through a cross-modal transformer architecture and modality-specific attention fusion for volatility and price movement prediction of stock market indices, gold price, currency exchange rates, and bond prices. We provide a cumulative of 24K data points for experimentation.
- MPCNet empirically outperforms other competitive deep learning approaches by 5-18% in this new task domain.

Financial assets	Impact of MPC announcements
Stock Prices (Large/ Small)	Indications of healthy, steadily growing economy increases stock prices Size of stock market - large *vs.* small, indicates set of all stocks *vs.* top performing stocks
Gold Price	Rise in inflation expectations raises prices of precious metals
Treasury bond yields (Short/Long-term)	Higher perceived risk of recession and rising interest rates leads to price increase Duration of bond term (short *vs.* long) indicates time expectation of interest rates hike
Currency Exchange Rate	Increase in employment and regulated inflation leads to appreciation in value

Table 1: Importance of MPC call analysis for financial forecasting.

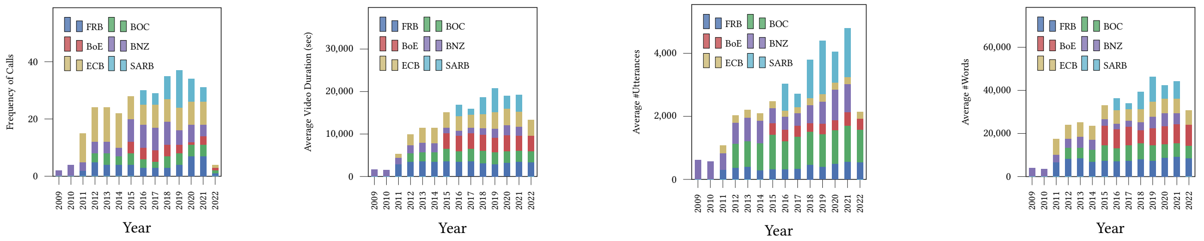
2 RELATED WORK

AI in Finance Traditional financial forecasting techniques have been applied in areas such as stock markets [3, 44], currency exchange markets [26, 56], and energy economics [6, 18]. Conventional financial models previously relied only on numerical features [36], which include discrete (ARIMA [3], GARCH [8], rolling regression [38]), continuous (Markov chain [24] and stochastic volatility [2]), and neural approaches [27, 32]. Efforts have since shifted towards utilizing textual data such as social media posts, news reports, web searches, etc. [49, 58]. These approaches limit their analyses to stock markets. [48] explored a multi-task setting for financial risk forecasting in stock markets using earnings calls. However, the multi-task setting is limited to simultaneous prediction of movement and volatility of a single target variable, and simultaneous prediction of multiple economic variables presents a new avenue for research in financial forecasting.

Monetary Policy Calls Previous research has shown that MPC calls provide key economic indicators that determine how the policy impacts the financial markets, and can improve financial predictions [9, 43]. Studies have also been carried out exclusively for MPC calls [17, 53], which show that monetary policy meeting minutes affect policy expectations, often exerting an even larger effect on financial markets than the release of the policy decisions. Furthermore, the Q&A portion of the press conference serves as a clarifier of the economic outlook, particularly during times of high macroeconomic uncertainty [17]. There is, however, a gap in leveraging neural predictive modeling using visual, verbal and vocal cues pertaining to MPC calls for financial forecasting.

Multimodality in Financial Forecasting Existing work in the financial realm utilize vocal and textual cues from earnings conference calls [41, 48], and mergers and acquisitions calls [47] for stock volatility prediction. Multimodal architectures that use these cues for financial predictions have seen significant improvements in their performances [48, 60]. However, the vision modality, which may offer important cues that correlate with the performance of

[2] https://github.com/monopoly-monitory-policy-calls/MONOPOLY



(a) MPC call frequency **(b) Video duration** **(c) Mean # utterances.** **(d) Mean # words.**

Figure 2: Year-wise statistics for each bank (FRB: Federal Reserve Bank of USA, BOC: Bank Of Canada, BoE: Bank of England, BNZ: Reserve Bank of New Zealand, ECB: European Central Bank, SARB: South African Reserve Bank).

financial markets [11] remains underexplored, which we seek to address with this work.

3 PROBLEM FORMULATION

We consider a monetary policy meeting χ which consists of three components: $\chi = [v; a; t]$. The sequence of textual utterances[3] $t = [t_1, t_2, \cdots, t_N]$ is extracted from the meeting transcript where t_i is the i^{th} utterance of the call and N is the maximum number of utterances in any call. Similarly, a is the sequence of corresponding audios for the textual utterances and is represented as $[a_1, a_2, \cdots, a_N]$. Finally, v corresponds to the sequence of the video frames corresponding to each audio segment, given by $[v_1, v_2, \cdots, v_N]$. Each utterance in a given call belongs to speaker $s \in \{governor, reporter\}$. Our goal is to forecast predictions for the set of six principle financial targets: $\mathbb{U}$ = {*Stock Index (Small), Stock Index (Large), Gold Price, Currency Exchange Rate, Long-term bond yield (10-years), Short-term bond yield (3-months)*}. We experiment simultaneously predicting all target variables using shared model parameters. For volatility prediction, we stack all computed volatility values $v^u_{[d,d+\tau]}, \forall u \in \mathbb{U}$ into an $|\mathbb{U}|$-dimensional target vector $\mathbf{v}_{[d,d+\tau]}$. For the movement prediction task, we similarly stack all computed movement labels $y^u_{[d,d+\tau]}, \forall u \in \mathbb{U}$ into an $|\mathbb{U}|$-dimensional target vector $\mathbf{y}_{[d,d+\tau]}$. We will now describe the two kinds of prediction tasks that we explore in this work i.e volatility and movement prediction.

Volatility: Following [28], we define volatility prediction as a regression problem. For a given target variable $u \in \mathbb{U}$ with price p_i on day i, the volatility is the natural log of the standard deviation of return prices r in a window of τ days, given as,

$$v^u_{[d,d+\tau]} = \ln\left(\sqrt{\frac{\sum_{i=d}^{d+\tau}(r_i - \bar{r})^2}{\tau}}\right), \ v \in \mathbb{R} \tag{1}$$

where $r_i = \frac{p_i - p_{i-1}}{p_{i-1}}$ is the return price on day i of the target m, and $\bar{r}$ is the average of these returns over a period of τ days.

Price movement Following [59], we define price movement $y_{[d,d+\tau]}$ over a period of τ days as a binary classification task. For a given target, whose price p can either rise or fall on a day $d + \tau$ compared to a previous day d , we formulate the classification task,

$$y^u_{[d,d+\tau]} = \begin{cases} 1, & p_{d+\tau} \geq p_d \\ 0, & p_d \geq p_{d+\tau} \end{cases} \tag{2}$$

Bank	Year Range	# of Data Samples
Federal Reserve	2011-22	3804
European Central Bank	2011-22	7416
Bank of England	2015-22	1728
Bank of Canada	2012-22	2808
Reserve Bank of New Zealand	2009-22	5040
South African Reserve Bank	2016-22	3384

Table 2: Data distribution of conference video files for each bank. Number of data samples corresponds to total data points in the Monopoly dataset corresponding to each bank.

4 MONOPOLY DATASET

Conference call transcripts and audios have been extensively studied in the past [41, 47]. However, there is no existing financial conference dataset that captures the visual modality. Therefore, we present the Monopoly dataset with videos, audio recordings and text transcripts corresponding to the monetary policy committee meetings conducted by the central banks of six major economies - United States, United Kingdom, European Union, Canada, New Zealand, and South Africa. To limit the scope, we ensured all audios and transcripts were in English, and had "Monetary Policy" mentioned in their titles.

4.1 Dataset Acquisition

We extract the conference call videos from the official websites of the respective central banks as mentioned in Table 5 in Appendix. We used BeautifulSoup[4] Python package to web scrape the dates, video links, and transcripts of the monetary policy calls, and download the MP4 videos and PDF transcripts using Urllib[5]. Textual components of the PDF were extracted using PDFPlumber[6] python library. Table 7 in Appendix points to the reference ticker names used in different financial markets to indicate the market prices for stock indices (large and small), gold price, currency exchange rate, long-term and short-term bond yields. We use the Bloomberg Terminal[7] to extract the time series of daily prices between Jan

[3] Due to higher complexity and noise of processing long length of videos, we segment at sentence level as opposed to the word level.

[4] https://www.crummy.com/software/BeautifulSoup/

[5] https://pypi.org/project/urllib3/

[6] https://pypi.org/project/pdfplumber/

[7] https://bba.bloomberg.net/

2000 to Mar 2022 corresponding to the six financial target for each conference call.

4.2 Dataset Statistics

Since conference calls started being reliably released post 2009, we filter and list all MPC calls between January 2009 and March 2022. These meetings are held 8 times in a year. A total of 464 MPC conference calls were downloaded. However, we discarded conference calls where text-audio-video alignment was not possible due to missing media or transcription files. The final dataset comprises of 340 conference calls of a combined duration of $15,729$ minutes with the average duration of the calls around 53 minutes. The scripted opening statement during the press conference is on average just shy of 10 minutes long, while the Q&A session usually lasts for about 44 minutes, with the governor answering an average of 22 questions and follow-ups. Table 2 shows the data distribution for conference calls originating from different banks. The mean number of audio utterances across the calls is 587.54 ± 38.32, with a maximum of 2462 utterances. Similarly, we observe varying lengths of conference calls with mean and maximum number of words as 6280 and 17,258 words, respectively. Table 8 in Appendix gives further descriptive dataset statistics. Looking at year-wise trends in Figure 2, we see that the availability of calls gets more consistent every year as more and more countries mandate public release of conference recordings. We also see a positive trend of progressively increase in all three modalities of the conference calls - total duration (visual), number of utterances (vocal), as well as the number of words (textual) each year. The dataset is split chronologically into a train, validation, and test set in the ratio of $70 : 10 : 20$, respectively, to ensures that future data is not used for forecasting past data.

5 METHODOLOGY

5.1 Multi-Modal Segmentation and Alignment

Given the three modalities v, a and t, it is essential to segment them into sequences such that they align and correspond with each other. To perform segmentation, we follow existing work [47] and use utterance-level embeddings, where we consider each sentence or phrase as an utterance. We perform forced alignment using the library Aeneas[8] to align the audio segments with textual utterance. Aeneas uses the Sakoe-Chiba Band Dynamic Time Warping (DTW) [46] forced alignment algorithm, which shows high discrimination between words. The Forced Alignment algorithm takes as input a text file divided into segments $t = [t_1, t_2, \cdots, t_N]$, an unfragmented audio file a, and returns a mapping which associates each text fragment $t_j \in t$ with a corresponding time-interval in the audio file, given as $\hat{a} = [a(\tau_s^1, \tau_e^1), a(\tau_s^2, \tau_e^2), \cdots, a(\tau_s^N, \tau_e^N)]$, where $\hat{a}_j = a(\tau_s^j, \tau_e^j)$ is the j-th audio segment between timestamps τ_s^j and τ_e^j. Video frames are already aligned to the audio, i.e for a given audio segment a_j with start and end times of τ_s^j and τ_e^j respectively, we obtain the corresponding video segment $v_j = [v_j^1, v_j^2, \cdots, v_j^N]$ as a sequence of frames, given as $v_j = [v(\eta\tau_s^j), v(\eta\tau_s^j + 1), \cdots, v(\eta\tau_e^j)]$, where $v(k)$ denotes the k-th frame of the full video, η is the frame rate (in fps), and $s < e$. We use audio sampling rate of 44kHz and video frame rate of 12 fps for audio and video time series.

5.2 Multi-Modal Feature Extraction

Textual Features: We compute the feature representation of each utterance using BERT [14], which has shown to be an effective pre-trained language-based model for extracting word-embeddings. We embed each text utterance $t_j \in [t_1, t_2, \cdots, t_M]$ as the arithmetic mean of all its word representations from BERT, and obtain a text encoding $k_j \in \mathbb{R}^{768}$, given as $x_T^j = \text{BERT}(t_j)$, $\forall j \in [1, N]$. We thus obtain a sequence of text embeddings $X_T = [x_T^1, x_T^2, \cdots, x_T^N]$.

Audio Features: To encode audio segments, we use wav2vec2 [4], which has shown shown significant potential for extracting audio features for speech language understanding tasks. We embed each audio utterance a_j as the arithmetic mean of the output representation from wav2vec2, to obtain an audio encoding $l_j \in \mathbb{R}^{768}$, given as $x_A^j = \text{wav2vec2}(\hat{a}_j)$, $\forall j \in [1, N]$. The sequence of audio embeddings is represented as $X_A = [x_A^1, x_A^2, \cdots, x_A^N]$.

Video Features: We encode the video frames using BEiT [5], which is a pre-trained bidirectional transformer based encoder for extracting image representations. BEiT has shown great promise for obtaining pre-trained representations for downstream vision tasks [23]. We embed each frame v_j^k in the video fragment v_j as the arithmetic mean of visual tokens representations of that frame. We then average over all the frames to obtain the aggregated encoding $x_V^j \in \mathbb{R}^{768}$ of the segment v_j, given as $x_V^j = \frac{1}{L}\sum_{k=1}^{L} \text{BEiT}(v_j^k)$, $\forall j \in [1, N]$, where L is the number of frames in the segment v_j. The sequence of video embeddings is represented as $X_V = [x_V^1, x_V^2, \cdots, x_V^N]$.

5.3 MPCNet: MPC Crossmodal Transformer

Due to the multimodal nature of the data, the model must learn the correlations and inter-dependencies between modalities. The model needs to accurately contrast visual, auditory, and textual information in order to characterize the speaker's affective state [13, 34, 52]. Hence, we leverage and build upon crossmodal transformers [54, 63], which have shown to be effective for learning fused multimodal representations through latent crossmodal adaptation. Let the set of available modalities be represented as $\mathbb{M} = \{V, A, T\}$, namely Video, Audio and Text respectively. The basic building block of the crossmodal transformer is the crossmodal attention module, which reinforces source modality α with target modality β using their respective locally-enriched feature sequences.

Locally-Aware Positional Encoding[47]: Given input sequence $X_\alpha \in \mathbb{R}^{L \times 768}$, where $\alpha \in \mathbb{M}$, we first pass this representation through a 1D temporal convolutional layer to capture the local sequence structure [54, 61]. This step produces a locally-aware sequence of features $\hat{X}_\alpha$, given as $\hat{X}_\alpha = \text{Conv1D}(X_\alpha)$, $\forall \alpha \in \mathbb{M}$. To enable the sequences to carry temporal information [54, 55], we augment positional embedding pos to locally-aware features $\hat{X}_\alpha$ to yield position enriched features $\widetilde{X}_\alpha = \hat{X}_\alpha + pos$, where pos is,

$$pos_{j,2l}, pos_{j,2l+1} = \sin\left(\frac{j}{10^{\frac{8l}{d}}}\right), \cos\left(\frac{j}{10^{\frac{8l}{d}}}\right) \tag{3}$$

$$\tag{4}$$

Crossmodal Attention[54]: For two modalities $\alpha, \beta \in \mathbb{M}$ where $\alpha \neq \beta$, the crossmodal attention layer fuses crossmodal information

[8] https://github.com/readbeyond/aeneas

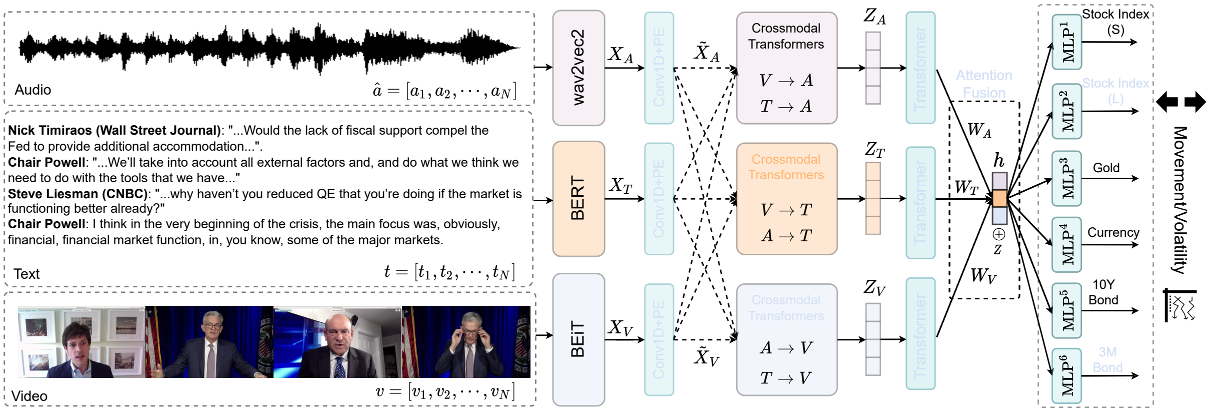

Nick Timiraos (Wall Street Journal): "...Would the lack of fiscal support compel the Fed to provide additional accommodation...".
Chair Powell: "...We'll take into account all external factors and, and do what we think we need to do with the tools that we have..."
Steve Liesman (CNBC): "...why haven't you reduced QE that you're doing if the market is functioning better already?"
Chair Powell: I think in the very beginning of the crisis, the main focus was, obviously, financial, financial market function, in, you know, some of the major markets.

Figure 3: We illustrate each building block in the architectural pipeline of MPCNet, starting with i) feature extraction ii) locally-aware position encoding iii) crossmodal transformer blocks iv) sentence-level transformers v) feature-fusion, and finally vi) target-specific MLPs for prediction.

through latent adaptation between α and β [54]. Given position-aware features $Z^{i-1}_{\alpha\to\beta}$ and Z^{i-1}_{α} at the $(i-1)^{\text{th}}$ transformer block, the intermediate latent adaption $\hat{Z}^{i}_{\alpha\to\beta}$ is computed as,

$$\tilde{Z}^{i-1}_{\alpha\to\beta} = \text{LN}(Z^{i-1}_{\alpha\to\beta}),\ \tilde{Z}^{i-1}_{\alpha} = \text{LN}(Z^{i-1}_{\alpha}) \tag{5}$$

$$\hat{Z}^{i}_{\alpha\to\beta} = \text{softmax}\left(\frac{\tilde{Z}^{i-1}_{\alpha\to\beta} W_q W_k^{\top} (\tilde{Z}^{i-1}_{\alpha})^{\top}}{\sqrt{d}}\right) \tilde{Z}^{i-1}_{\alpha} W_v + \tilde{Z}^{i-1}_{\alpha\to\beta} \tag{6}$$

$W_{(\cdot)}$ are learnable weight matrices, and d is the feature dimension, LN means layer-norm, and $Z^{0}_{\alpha\to\beta} = \widetilde{X}_{\beta}$. The intermediate latent adaption $\hat{Z}^{i}_{\alpha\to\beta}$ is then passed through a feedforward (FF) layer to yield $Z^{i}_{\alpha\to\beta}$ as $Z^{i}_{\alpha\to\beta} = \text{FF}(\text{LN}(\hat{Z}^{i}_{\alpha\to\beta})) + \text{LN}(\hat{Z}^{i}_{\alpha\to\beta})$.

Sentence-Level Transformer[54]: We concatenate $Z_{\alpha\to\beta}$ from the crossmodal transformers sharing the same target modality $\beta \in \mathbb{M}$ to yield modality specific representations $Z_{\alpha}, \forall\alpha \in \mathbb{M}$, given as $Z_V = [Z_{T\to V}; Z_{A\to V}]$, $Z_A = [Z_{T\to A}; Z_{V\to A}]$, $Z_T = [Z_{V\to T}; Z_{A\to T}]$.

Next, these hidden states are passed through self-attention transformers [54, 55] to collect temporal information. The temporal encodings are then concatenated and passed through a feed forward layer to yield the ensembled temporal representation Z.

Modality Specific Attention-Fusion We propose an additional attention fusion mechanism to capture the importance of a specific target modality representation Z_{α} with respect to sibling representations Z_{β} ($\alpha \neq \beta$). We first compute attention weights $W_{\alpha}, \forall\alpha \in \mathbb{M}$ for video, audio and textual representations respectively, given as,

$$W_{\alpha} = \frac{\widetilde{W}_{\alpha}}{\sum_{\alpha' \in \mathbb{M}} \widetilde{W}'_{\alpha}}, \quad \text{where } \widetilde{W}_{\alpha} = \text{softmax}(\hat{W}_{\alpha} Z_{\alpha} + \hat{b}_{\alpha}) \tag{7}$$

where $\hat{W}_{\alpha}$ and $\hat{b}_{\alpha}$ are learnable parameters, and $\alpha \in \mathbb{M}$. We then fuse the attention video, audio and textual features by multiplying the computed weights with their corresponded feature representations to yield the fused temporal representation $Z_{\text{fused}} = \sum_{\alpha\in\mathbb{M}} W_{\alpha} Z_{\alpha}$

Final Network and Prediction: Finally, we combine the ensembled temporal representation Z with the fused temporal representation Z_{fused} by using a feed forward layer with a residual block to yield the final hidden representation h, given as $h = FF(Z_{\text{fused}}) + Z$. The final hidden representation is then passed through $|\mathbb{U}|$ multilayer perceptrons (MLPs) to yield the prediction $y^u, \forall u \in \mathbb{U}$ as $y^u = \sigma(\text{MLP}^u(h))$, where σ represents the final activation function. We use a linear activation for volatility prediction and sigmoid for price movement, respectively. We use Mean Squared Error (MSE) and Binary Cross-Entropy (BCE) for these tasks, respectively.

6 EXPERIMENTS

Baselines: We compare MPCNet against several modern and traditional baselines across varied domains and modalities as follows:

6.0.1 *Price-based Baselines.* : Utilizing historical price exclusively.

- **HistPrice**: Following [16], we use ARIMA model to perform regression/classification on past 30-days time series.
- **P-SVM** [12]: We apply Support Vector Regression (SVR) and Classifiers (SVC) on 30-days historical price data for volatility and price movement prediction, respectively.
- **P-LSTM** [62]: We use LSTM model to extract predictive patterns from 30-days historical price time-series.

6.0.2 *Multimodal Baselines.* : We present contemporary multimodal methods that utilize visual, vocal, and verbal cues.

- **MLP**: A simple multi-layer perceptron where multimodal features are averaged out along the time series and concatenated before the final prediction layer.
- **LSTM [40]**: Multimodal time series are input to individual LSTMs and averaged before final prediction.
- **MMIM [22]**: Uses LSTMs to encode the video and audio sequence, and BERT for text. The encoded features are passed through a fusion layer for maximizing mutual information between unimodal sequences before prediction.

	Model	Stock Index (Small)				Stock Index (Large)				Currency Exchange Rate			
		$\text{MSE}_1\downarrow$	$\text{MSE}_3\downarrow$	$\text{MSE}_7\downarrow$	$\text{MSE}_{15}\downarrow$	$\text{MSE}_1\downarrow$	$\text{MSE}_3\downarrow$	$\text{MSE}_7\downarrow$	$\text{MSE}_{15}\downarrow$	$\text{MSE}_1\downarrow$	$\text{MSE}_3\downarrow$	$\text{MSE}_7\downarrow$	$\text{MSE}_{15}\downarrow$
Baselines	HistPrice	2.486	2.234	1.880	1.664	3.397	3.316	2.934	2.972	2.709	3.187	3.127	3.291
	P-SVM [12]	2.489	2.220	1.915	1.753	2.568	2.921	1.971	2.012	2.104	2.534	1.921	2.231
	P-LSTM [62]	2.421	2.217	1.845	1.731	2.128	2.194	2.108	1.456	1.424	1.867	1.015	1.569
	MLP	2.524	2.214	1.899	1.680	1.469	1.597	0.937	0.981	1.060	1.441	0.802	1.159
	LSTM [40]	2.290	2.210	1.750	1.680	1.346	1.304	0.724	0.779	1.219	1.296	0.762	0.558
	MMIM [22]	2.290	*2.092*	1.779	1.598	1.287	1.133	0.718	0.622	**0.975***	1.081	*0.500*	0.510
	MDRM [41]	**2.065**	2.511	*1.748*	*1.597*	*1.281*	1.578	*0.683*	0.612	1.183	1.627	0.769	0.512
	HTML [60]	2.296	2.133	1.771	1.611	1.302	*1.127*	0.766	*0.609*	*0.988*	1.118	0.588	0.498
	MULT [54]	*2.073*	2.179	1.768	1.605	1.288	1.133	**0.672***	0.742	1.022	*1.018*	0.549	*0.497*
Ablation	MPCNet (T)	2.599	2.390	1.931	2.278	1.906	1.613	1.122	1.262	1.666	1.943	*1.140*	1.801
	MPCNet (A)	2.345	2.457	*1.770*	2.151	1.732	1.614	1.221	*0.724*	1.507	1.963	1.289	1.791
	MPCNet (V)	2.532	2.285	2.108	2.023	1.904	1.617	1.223	1.247	2.273	1.964	1.746	1.511
	MPCNet (T+A)	2.423	*2.221*	2.135	1.956	1.564	1.637	1.456	1.111	1.234	2.144	1.967	1.578
	MPCNet (A+V)	2.280	2.413	2.026	*1.680*	1.857	*1.572*	1.697	0.864	1.621	*1.904*	1.419	1.463
	MPCNet (V+T)	*2.257*	2.321	2.002	2.108	*1.477*	1.596	*1.195*	1.398	1.087	2.017	1.819	*1.407*
	MPCNet (V+A+T) (Ours)	2.233	**2.089***	**1.732***	**1.594***	**1.269***	**1.046***	0.806	**0.607**	1.176	**1.001**	**0.469***	**0.470***

(a) Stock Indices and Currency Exchange Rate

	Model	Gold Price				10-Year Bond Yield				3-Month Bond Yield			
		$\text{MSE}_1\downarrow$	$\text{MSE}_3\downarrow$	$\text{MSE}_7\downarrow$	$\text{MSE}_{15}\downarrow$	$\text{MSE}_1\downarrow$	$\text{MSE}_3\downarrow$	$\text{MSE}_7\downarrow$	$\text{MSE}_{15}\downarrow$	$\text{MSE}_1\downarrow$	$\text{MSE}_3\downarrow$	$\text{MSE}_7\downarrow$	$\text{MSE}_{15}\downarrow$
Baselines	HistPrice	3.193	3.039	2.675	2.683	4.132	4.020	3.472	3.334	3.899	3.665	3.063	2.913
	P-SVM [12]	2.568	2.543	1.967	2.104	3.212	3.589	2.986	3.141	3.235	3.143	2.922	2.874
	P-LSTM [62]	1.965	1.998	1.043	1.764	2.212	1.699	2.340	1.453	3.433	2.909	2.678	2.477
	MLP	1.431	1.654	0.904	0.955	1.811	1.743	1.288	1.382	2.582	2.523	2.239	2.231
	LSTM [40]	1.472	1.484	0.703	0.508	1.735	1.801	1.169	1.235	2.421	2.439	2.044	2.013
	MMIM [22]	*1.292*	1.292	*0.565*	0.486	*1.698*	*1.604*	*1.080*	*1.053*	2.345	2.392	1.977	*1.902*
	MDRM [41]	1.436	1.843	0.710	*0.483*	1.729	1.699	1.126	1.223	2.406	2.622	2.096	1.993
	HTML [60]	**1.277***	*1.291*	0.589	0.524	**1.685**	1.612	1.103	1.149	*2.342*	*2.356*	*1.962*	1.998
	MULT [54]	1.314	1.335	0.579	0.503	2.122	1.837	1.104	**1.037***	**1.174***	2.515	1.973	1.903
Ablation	MPCNet (T)	1.967	1.859	1.122	1.750	1.977	1.928	2.067	1.614	2.774	2.723	2.654	2.602
	MPCNet (A)	*1.573*	1.484	1.617	1.803	2.279	1.940	1.965	1.513	2.754	3.242	2.726	2.536
	MPCNet (V)	2.136	2.028	1.586	1.158	2.318	1.969	1.576	1.674	2.857	2.740	2.630	2.616
	MPCNet (T+A)	1.798	1.567	0.985	1.678	2.067	1.956	1.944	1.865	2.759	2.699	*2.345*	2.613
	MPCNet (A+V)	1.752	*1.403*	1.245	*0.959*	1.996	1.903	1.897	1.700	*2.750*	*2.632*	2.538	*2.527*
	MPCNet (V+T)	1.681	1.959	*0.864*	1.428	1.756	1.874	1.511	1.366	3.135	2.678	2.457	2.564
	MPCNet (V+A+T) (Ours)	1.342	**1.275***	**0.562***	**0.477***	1.767	**1.602***	**0.979***	1.142	2.431	**2.319***	**1.948***	**1.901***

(b) Gold Prices, Long-term (10-Years) and Short-term (3-Months) Bonds

Table 3: Performance comparison with baselines and ablations for volatility prediction in terms of MSE τ-days after the call ($\tau \in \{1, 3, 7, 15\}$). (T: Text, V: Video, A: Audio). Bold denotes best performance performance. Light cyan shows second-best performance. Results are averaged over 5 independent runs. * indicates that the result is statistically significant with respect to state-of-the-art based on the Wilcoxon's signed rank test with $p < 0.001$. Our proposed approach outperforms price-based and multimodal baselines.

- **MDRM [41]**: BiLSTM layers encode unimodal sequences, which are then fused together using another layer of BiLSTM to extract multimodal inter-dependencies.
- **HTML [60]**: HTML is a transformer based architecture that takes fuses multimodal feature representations before passing through Transformer layers for prediction.
- **MulT [54]**: Uses transformer encoders to align language, facial gestures, and acoustic sequences with variable sampling rates and long-range dependencies.

Experiment Settings: MPCNet uses a hidden dimension $H = 512$, dropout $\delta = 0.1$, number of attention heads $n_h = 2$, and number of transformer blocks $n_b = 2$. We use a learning rate (lr) of $1e^{-3}$ for regression, and $1e^{-4}$ for classification. We use PyTorch for all models, and optimize MPCNet using AdamW optimizer for 30 epochs and apply early stopping with a patience of 10 on a Tesla K80 GPU. We summarize the range of hyperparameters in Sec-E of Appendix.
Evaluation Metrics: Similar to prior work [41, 60], we evaluate predicted volatility using the mean squared error (MSE) and the price movement classification task using F1 score, for $\tau \in \{1, 3, 7, 15\}$.

7 RESULTS

Performance Comparison: Tables 3 and 4 show the comparative results for the volatility and price prediction tasks, respectively. We observe that baselines that use historical price alone significantly under perform across all settings. Simple models like MLP and LSTM are disadvantaged as they require feature aggregation through averaging over long sequences of time series. Sophisticated LSTM models such as MMIM and MDRM struggle on both tasks due to their inability to capture long-range dependencies in hour long video calls with multiple dialogues. Combining multimodal context from the visual, vocal and verbal cues using a transformer encoder (as done in HTML and MuLT) helps improve performance across different settings. Our proposed model achieves significantly better performance across both tasks for multiple financial targets. MPCNet's ability to model the inter-dependencies between the pairs of modalities using cross-model attention and modality-specific attention fusion contributes towards its outperformance compared to contemporary multimodal methods. Moreover, MPCNet performs attention fusion using weights for pairs of the modalities to determine the mutual importance of each modality which helps it improve over MuLT baseline. However, it can also be observed that there is ample room for improvement for both volatility and price movement prediction. We attribute this to the inherent difficulty of task and motivate further research by discussing current shortcomings through error analysis in Sec-8.
Ablation: Impact of Multimodality: The ablation results of the proposed MPCNet model in Tables 3 and 4 strongly suggest the potency of multimodal features over unimodal counterparts, for both tasks, across all financial targets. We observe significant gains due to addition of aligned video features in the MPCNet model. We attribute this to the presence of additional behavioral cues such as facial expressions and body language, aligned with the call transcripts and audio signals through attention mechanisms in the temporal domain. In order to validate the importance of combining visual, vocal and verbal cues, we conduct additional ablation experiments for MDRM, HTML, and MuLT baselines with varying input modalities. Figure 4 shows that blending video features (V) with text(T) and audio(A) leads to improvements over the best bimodal model (T+A), evaluated in terms of time-averaged MSE and F1 scores for MPCNet. We see a similar trend for HTML, MDRM, and MuLT as depicted by figures 9, 10, and 11 in the Appendix, respectively. Moreover, we see that the addition of video (V) modality to each of *A, T, A+T* settings shows favourable gains. This provides strong empirical evidence in support of multimodal fusion of visual, vocal and verbal modalities for financial prediction tasks on MPC calls.
Impact of Call Length: We probe MPCNet's sensitivity with respect to the input call length by feeding only the first n utterances of the call to the model. As shown in Figure 5, we see major performance improvements with increasing call length, and achieve best performance on incorporating the full conference call. These observations suggest that the Q&A session is substantially beneficial than just the initial speech by the governor, as the Q&A provides an opportunity to analyze non-verbal cues and answers to questions are not rehearsed beforehand. Our observations reinforce prior studies which have shown the importance of Q&A sessions, which serve as a clarifier of the overall economic outlook [17].

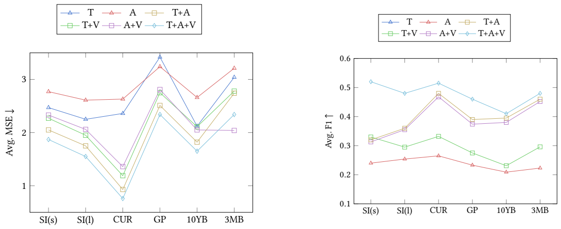
	Model	Stock Index (Small)				Stock Index (Large)				Currency Exchange Rate			
		$F1_1$↑	$F1_3$↑	$F1_7$↑	$F1_{15}$↑	$F1_1$↑	$F1_3$↑	$F1_7$↑	$F1_{15}$↑	$F1_1$↑	$F1_3$↑	$F1_7$↑	$F1_{15}$↑
Baselines	HistPrice	0.390	0.470	0.400	0.420	0.430	0.430	0.410	0.420	0.190	0.260	0.210	0.230
	P-SVM [12]	0.400	0.480	0.340	0.530	*0.433*	0.490	0.338	0.500	0.190	0.270	0.190	0.370
	P-LSTM [62]	0.410	0.473	0.291	0.546	0.399	0.391	0.421	0.442	0.123	0.232	0.165	0.341
	MLP	0.349	0.435	0.209	0.539	0.267	0.319	0.331	0.351	0.101	0.201	0.124	0.311
	LSTM [40]	0.449	*0.435*	0.269	0.527	0.414	0.596	0.371	0.432	0.137	0.229	0.199	0.369
	MMIM [22]	0.435	**0.653***	0.302	0.605	0.392	**0.631**	0.329	0.601	0.296	0.217	0.142	0.385
	MDRM [41]	0.449	0.419	0.462	0.355	0.409	0.392	0.494	0.324	0.177	0.161	*0.379*	0.152
	HTML [60]	0.490	*0.645*	0.458	0.541	0.431	0.504	0.557	0.482	0.484	0.531	0.298	**0.626***
	MULT [54]	*0.491*	0.630	*0.536*	*0.629*	*0.443*	*0.625*	*0.572*	*0.612*	*0.499*	*0.547*	**0.473***	*0.521*
Ablation	MPCNet (T)	0.393	0.423	0.241	0.263	0.361	0.304	0.419	0.396	0.332	0.215	0.252	0.378
	MPCNet (A)	0.288	0.233	0.182	0.365	0.211	0.315	0.397	0.435	0.410	0.283	0.111	0.331
	MPCNet (V)	*0.437*	0.522	0.340	*0.497*	0.335	0.304	*0.464*	0.443	0.438	0.148	0.254	0.412
	MPCNet (T+A)	*0.437*	0.569	0.289	0.489	0.367	0.312	0.422	0.471	0.404	0.245	0.392	0.466
	MPCNet (A+V)	0.415	0.565	0.290	0.465	0.388	0.321	0.455	*0.463*	0.434	0.186	*0.374*	*0.511*
	MPCNet (V+T)	0.406	*0.573*	*0.342*	0.469	0.359	*0.326*	0.458	0.405	*0.450*	*0.295*	0.350	0.336
	MPCNet (V+A+T) (Ours)	**0.501***	0.590	**0.565***	**0.638***	**0.460***	0.590	**0.559***	**0.620***	***0.520****	***0.570****	0.329	0.450

(a) Stock Indices and Currency Exchange Rate

	Model	Gold				10-Year Bond Yield				3-Month Bond Yield			
		$F1_1$↑	$F1_3$↑	$F1_7$↑	$F1_{15}$↑	$F1_1$↑	$F1_3$↑	$F1_7$↑	$F1_{15}$↑	$F1_1$↑	$F1_3$↑	$F1_7$↑	$F1_{15}$↑
Baselines	HistPrice	0.360	0.390	0.350	0.400	0.31	0.290	0.220	0.390	0.220	0.160	0.340	0.330
	P-SVM [12]	0.390	0.420	0.370	0.380	0.34	0.310	0.33	0.33	0.370	0.220	0.310	0.390
	P-LSTM [62]	0.365	0.352	0.371	0.346	0.32	0.291	0.342	0.258	0.377	0.234	0.332	0.314
	MLP	0.243	0.215	0.288	0.315	0.244	0.299	0.234	0.174	0.332	0.157	0.248	0.394
	LSTM [40]	0.361	0.337	0.304	0.345	0.364	0.311	0.255	*0.394*	0.381	0.168	0.382	0.444
	MMIM [22]	0.209	0.508	0.412	0.318	0.411	0.318	0.345	0.138	0.417	0.306	0.417	0.379
	MDRM [41]	*0.434*	0.383	0.214	0.317	0.287	0.242	0.314	0.149	0.346	0.198	**0.478***	*0.505*
	HTML [60]	*0.441*	*0.654*	0.379	0.526	*0.529*	0.278	0.466	0.389	0.424	*0.314*	0.397	0.450
	MULT [54]	0.329	0.590	**0.454**	*0.533*	**0.534**	**0.364***	*0.485*	*0.400*	*0.428*	0.171	*0.466*	0.493
Ablation	MPCNet (T)	0.341	0.317	0.423	0.492	0.242	0.343	0.155	0.592	0.117	0.437	0.310	0.293
	MPCNet (A)	0.292	0.121	0.119	0.589	0.088	0.157	0.186	0.489	0.252	0.386	0.317	0.314
	MPCNet (V)	0.239	0.414	*0.519*	0.595	0.373	0.436	*0.542*	0.610	*0.503*	0.520	0.314	*0.375*
	MPCNet (T+A)	0.414	*0.483*	0.503	*0.616*	0.322	0.434	0.529	0.593	0.476	*0.545*	*0.323*	0.312
	MPCNet (A+V)	*0.423*	0.445	0.414	0.607	0.372	0.416	0.449	*0.617*	*0.503*	0.510	0.309	0.369
	MPCNet (V+T)	0.420	0.472	0.517	0.565	0.471	*0.454*	0.500	0.585	0.485	0.542	0.315	0.347
	MPCNet (V+A+T) (Ours)	**0.444***	**0.668***	*0.413*	**0.637***	0.386	*0.327*	**0.560***	**0.625***	**0.493***	**0.556***	0.374	**0.537***

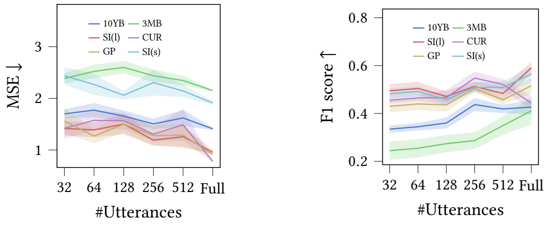
(b) Gold Prices, Long-term (10-Years) and Short-term (3-Months) Bonds

Table 4: Performance comparison with baselines and ablations for price prediction in terms of F1 score τ-days after the call ($\tau \in \{1, 3, 7, 15\}$). (T: Text, V: Video, A: Audio) Bold denotes best performance performance. Light cyan shows second-best performance. Results are averaged over 5 independent runs. * indicates that the result is statistically significant with respect to state-of-the-art based on the Wilcoxon's signed rank test with $p < 0.001$. Our proposed approach outperforms price-based and multimodal baselines.



(a) Volatility **(b) Movement**

Figure 4: Ablation analysis of modalities in MPCNet for (a) Volatility and (b) Price Movement prediction, averaged over $\tau = \{1, 3, 7, 15\}$. SI(s): Stock Index (Small), SI(l): Stock Index (large), CUR: Currency Exchange Rate, GP: Gold Price, 3MB: 3-Month Bond Yield, 10YB: 10-Year Bond Yield. Addition of video (V) modality to each of *A, T, A+T* settings shows favourable gains (increase in F1 and decrease in MSE).



(a) Volatility **(b) Movement**

Figure 5: Performance variation with increasing input call lengths (#utterances) on (a) Volatility and (b) Movement prediction. The results are averaged over $\tau = \{1, 3, 7, 15\}$. Performance improves with increasing call length (reduced MSE and increase in F1), with best results on full conference call.

Performance Drift over Time: Results in Table 3 and Figure 6 show that multimodal models exhibit greater uncertainty in the short term after the MPC call. However, there is a gradual decay in gains of multimodal models for volatility prediction as we move ahead in time after the conference call. This trend is not pronounced for price movement prediction which remains consistent throughout as observed from Table 4. Short-term stock volatility prediction is more complex due to the erratic price fluctuations after a MPC call. We attribute the saturation in the volatility prediction performance to the dilution of the market reaction to the MPC calls as we "drift" away from them. These price fluctuations settle as more time elapses, similar to the phenomenon of PEAD (Post Earnings Announcement Drift) [7, 45].

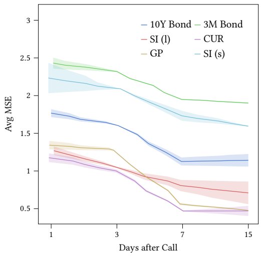
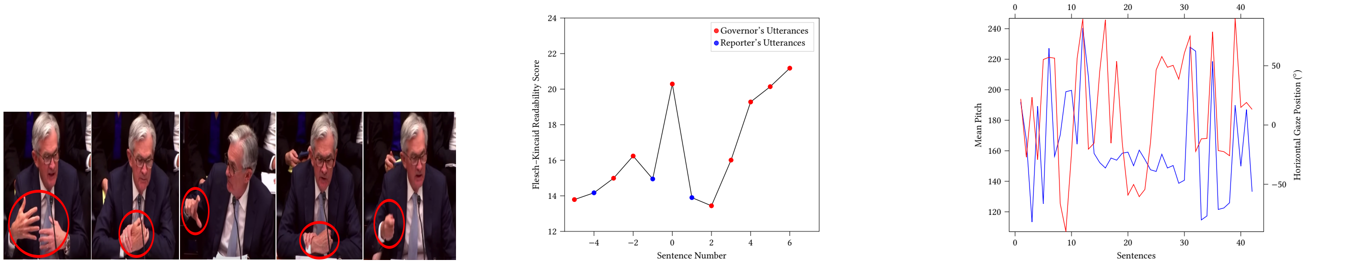


Figure 6: Drift in predicted stock volatility over time. The line graph represents the mean MSE of MPCNet SI(s): Stock Index (Small), SI(l): Stock Index (large), CUR: Currency Exchange Rate, GP: Gold Price. As time increases, the MSE decreases due to the PEAD phenomenon[7].

8 QUALITATIVE ANALYSIS

Video 1: Federal Reserve Meeting (2020): Following the MPC call, the SP500 suffered a significant drop within the next 20 days. Studying the call's video frames aligned with text transcripts, we notice in Figure 7a that when asked about their plans on interest rate during the Q&A session, the governor's speech had sudden fillers words along with animated hand gestures. Past research [39] suggests that increased use of filler words, rapid hand movements, and a closed body posture with hands crossed interlocked tightly may indicate a lack of confidence in the speaker. It was later ascertained that the Federal Reserve convened an emergency meeting a week later to announce interest rate cut of 0.50%. We observe how MPCNet successfully predicts the decrease in price of stock index and increase in gold prices for all choices of τ while it's unimodal (A,T,V) and bimodal (text-audio) counterparts fail to do the same each time. Though the text reveals no lack of confidence, the *combination of aligned audio-visual cues* likely allows the model to make a *successful prediction.*



(a) Video-1: Chair of the federal reserve exhibits closed body language, frequent interlocking of hands and enhanced use of filler words during Q&A session when asked about rising interest rates. Past research [1, 50] suggests these non-verbal cues indicate of lack of confidence.

(b) Video-2: Flesch–Kincaid Readability score of the utterances. The governor's utterances becomes more elongated and difficult to passage when questioned regarding employment reduction. Reporter's sentences are simple and easy to comprehend.

(c) Video-3: Erratic mean pitch of the governor's audio clips (—-) and rapid changes in horizontal gaze position (—-). Randomness in non-verbal cues adds noise, affecting predictions.

Figure 7: Qualitative Analysis

Video 2: European Central Bank (2016): Ten days post monetary policy conference, long-term and short-term bond saw an increase in volatility by 15-25%, respectively. However, the prices of long-term bond yield saw a downward trend contrary to the short-term bond yields. The meeting involved the governor mentioning concerns about trade disruptions and employment reduction due to 'Brexit'. We notice that this call in specific was longer than previous others. Anecdotally, longer conferences are linked with turbulent economic conditions as more time is spent clarifying journalist questions. We also observe enhanced complexity of text readability due to dense technical discussion in Q&A dialogues (Figure 7b). Transformer based models such as HTML, MuLT, and MPCNet were able to capture linguistic complexity and long-range dependencies. Here, we observe that the above three strategies correctly make correct predictions.

Video 3: South African Reserve Bank (2022): We now analyze this MPC call as an *error analysis* where `MPCNet` predicts incorrectly. Here, the price-based LSTM model gains a profit by correctly predicting a 9-12% increase in the currency exchange (ZAR USD rate) for $\tau = 3, 7, 15$. On carefully analyzing the contents of the conference call, we notice (Figure 7c) that the governor took a sudden hawkish stance on inflation due to oil crisis propelled by the Ukraine war and economic sanctions. Moreover, observing the visual and vocal cues, we find a great deal of variance in the mean audio pitch and speaker's erratic eye gaze. We attribute the erroneous performance to the potential overfitting of the model as well as unique information about world knowledge present at test time not seen before in the training set. We believe that future research in combining knowledge from *alternate sources such as news and social media can benefit prediction performance.*

9 ETHICAL CONSIDERATIONS AND LIMITATIONS

Examining a speaker's tone and speech in conference calls is a well-studied task in past literature [41, 60]. Our work focuses on video conference calls for which government institutions and financial regulatory bodies publicly release call videos, transcripts and audio recordings. The conference call and price data used in our study is open source. We do not collect any personalized data or violate any privacy laws in using, storing or releasing the MPC conference calls data for financial analysis.

Limitations: We acknowledge the presence of gender bias in our study, given the imbalance in the gender ratio of speakers of the calls. We also acknowledge the demographic bias in our study as the central banks studied in our work are restricted to certain geographies and may not directly generalize for other countries. We also limit our study to English-only calls, motivating further studies on other multilingual conference calls.

Potential risks: Our contributions are meant as an exploratory research in the financial domain and no part of the work should be treated as financial advice. All financial investments decisions are subject to market risk and should be made after extensive testing. Practitioners should check for various biases (demographic, gender, modeling, randomness) before attempting to use the provided code/data/methods for real-world purposes.

10 CONCLUSION AND FUTURE WORK

We present a dataset of Monetary Policy Conference video calls to predict financial risk and price movement. We also present `MPCNet`, a strong benchmark model that uses cross-modal transformer blocks and modality-specific attention fusion on input time series for financial forecasting on MPC calls. We further analyze the benefits of each modality, evaluate the effect of multi-task setting for joint prediction of financial assets, examine biases due to dataset distribution, and effect of non-verbal behavioral cues extracted from spontaneous Q&A session. We motivate future work to explore several interesting direction including but not limited to conversational dialogue modeling of Q&A sessions, fine-grained multimodal emotion recognition, gaits and posture analysis to identify non-verbal behavioural cues, augmenting video data with external knowledge graphs, etc.

REFERENCES

[1] Mohamed Abouelenien, Rada Mihalcea, and Mihai Burzo. 2016. Analyzing thermal and visual clues of deception for a non-contact deception detection approach. In *Proceedings of the 9th ACM International Conference on PErvasive Technologies Related to Assistive Environments*. 1–4.

[2] L Andersen. 2008. Simple and efficient simulation of the Heston stochastic volatility model. *Journal of Computational Finance* 11 (03 2008). https://doi.org/10.21314/JCF.2008.189

[3] Adebiyi A Ariyo, Adewumi O Adewumi, and Charles K Ayo. 2014. Stock price prediction using the ARIMA model. In *2014 UKSim-AMSS 16th International Conference on Computer Modelling and Simulation*. IEEE, 106–112.

[4] Alexei Baevski, Yuhao Zhou, Abdelrahman Mohamed, and Michael Auli. 2020. wav2vec 2.0: A Framework for Self-Supervised Learning of Speech Representations. In *Advances in Neural Information Processing Systems*, H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin (Eds.), Vol. 33. Curran Associates, Inc., 12449–12460. https://proceedings.neurips.cc/paper/2020/file/92d1e1eb1cd6f9fba3227870bb6d7f07-Paper.pdf

[5] Hangbo Bao, Li Dong, Songhao Piao, and Furu Wei. 2022. BEiT: BERT Pre-Training of Image Transformers. In *International Conference on Learning Representations*. https://openreview.net/forum?id=p-BhZSz59o4

[6] PMR Bento, JAN Pombo, MRA Calado, and SJPS Mariano. 2018. A bat optimized neural network and wavelet transform approach for short-term price forecasting. *Applied energy* 210 (2018), 88–97.

[7] Victor L Bernard and Jacob K Thomas. 1989. Post-earnings-announcement drift: delayed price response or risk premium? *Journal of Accounting research* 27 (1989), 1–36.

[8] Tim Bollerslev. 1986. Generalized autoregressive conditional heteroskedasticity. *Journal of econometrics* 31, 3 (1986), 307–327.

[9] Ellyn Boukus and Joshua V Rosenberg. 2006. The information content of FOMC minutes. *Available at SSRN 922312* (2006).

[10] Yong Cai, Santiago Camara, and Nicholas Capel. 2021. It's not always about the money, sometimes it's about sending a message: Evidence of Informational Content in Monetary Policy Announcements. *arXiv preprint arXiv:2111.06365* (2021).

[11] Longbing Cao. 2022. AI in Finance: Challenges, Techniques, and Opportunities. *ACM Computing Surveys (CSUR)* 55, 3 (2022), 1–38.

[12] Sotirios P Chatzis, Vassilis Siakoulis, Anastasios Petropoulos, Evangelos Stavroulakis, and Nikos Vlachogiannakis. 2018. Forecasting stock market crisis events using deep and statistical machine learning techniques. *Expert systems with applications* 112 (2018), 353–371.

[13] Chen Chen, Zuxuan Wu, and Yu-Gang Jiang. 2016. Emotion in context: Deep semantic feature fusion for video emotion recognition. In *Proceedings of the 24th ACM international conference on Multimedia*. 127–131.

[14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Association for Computational Linguistics, Minneapolis, Minnesota, 4171–4186. https://doi.org/10.18653/v1/N19-1423

[15] Theu Dinh, Stéphane Goutte, Duc Khuong Nguyen, and Thomas Walther. 2022. Economic drivers of volatility and correlation in precious metal markets. *Journal of Commodity Markets* (2022), 100242.

[16] Ning Du and David V Budescu. 2007. Does past volatility affect investors' price forecasts and confidence judgements? *International Journal of Forecasting* 23, 3 (2007), 497–511.

[17] Michael Ehrmann and Marcel Fratzscher. 2007. Explaining monetary policy in press conferences. (2007).

[18] Hamed Ghoddusi, Germán G Creamer, and Nima Rafizadeh. 2019. Machine learning in energy economics and finance: A review. *Energy Economics* 81 (2019), 709–727.

[19] Joumana Ghosn and Yoshua Bengio. 1996. Multi-task learning for stock selection. *Advances in neural information processing systems* 9 (1996).

[20] Roberto Gómez-Cram and Marco Grotteria. 2022. Real-time price discovery via verbal communication: Method and application to Fedspeak. *Journal of Financial Economics* 143, 3 (2022), 993–1025.

[21] Yuriy Gorodnichenko, Tho Pham, and Oleksandr Talavera. 2021. *The voice of monetary policy*. Technical Report. National Bureau of Economic Research.

[22] Wei Han, Hui Chen, and Soujanya Poria. 2021. Improving Multimodal Fusion with Hierarchical Mutual Information Maximization for Multimodal Sentiment Analysis. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 9180–9192. https://doi.org/10.18653/v1/2021.emnlp-main.723

[23] Ali Hatamizadeh, Yucheng Tang, Vishwesh Nath, Dong Yang, Andriy Myronenko, Bennett Landman, Holger R Roth, and Daguang Xu. 2022. Unetr: Transformers for 3d medical image segmentation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*. 574–584.

[24] Eric Jacquier, Nicholas G Polson, and Peter E Rossi. 2002. Bayesian analysis of stochastic volatility models. *Journal of Business & Economic Statistics* 20, 1 (2002), 69–87.

[25] Weiwei Jiang. 2021. Applications of deep learning in stock market prediction: recent progress. *Expert Systems with Applications* 184 (2021), 115537.

[26] Joarder Kamruzzaman and Ruhul A Sarker. 2003. Forecasting of currency exchange rates using ANN: A case study. In *International Conference on Neural Networks and Signal Processing, 2003. Proceedings of the 2003*, Vol. 1. IEEE, 793–797.

[27] Raehyun Kim, Chan Ho So, Minbyul Jeong, Sanghoon Lee, Jinkyu Kim, and Jaewoo Kang. 2019. Hats: A hierarchical graph attention network for stock movement prediction. *arXiv preprint arXiv:1908.07999* (2019).

[28] Shimon Kogan, Dimitry Levin, Bryan R Routledge, Jacob S Sagi, and Noah A Smith. 2009. Predicting risk from financial reports with regression. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. 272–280.

[29] Katharina Lewellen. 2006. Financing decisions when managers are risk averse. *Journal of Financial Economics* 82, 3 (2006), 551–589.

[30] Yelin Li, Junjie Wu, and Hui Bu. 2016. When quantitative trading meets machine learning: A pilot survey. In *2016 13th International Conference on Service Systems and Service Management (ICSSSM)*. IEEE, 1–6.

[31] Yue Liu, Haoyuan Feng, and Kun Guo. 2021. The Dynamic Relationship between Macroeconomy and Stock Market in China: Evidence from Bayesian Network. *Complexity* 2021 (2021).

[32] Rui Luo, Weinan Zhang, Xiaojun Xu, and Jun Wang. 2018. A neural stochastic volatility model. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 32.

[33] Alexis Marchal. 2021. Risk and Returns Around FOMC Press Conferences: A Novel Perspective from Computer Vision. In *Proceedings of SAI Intelligent Systems Conference*. Springer, 724–735.

[34] Claude Montacié and Marie-José Caraty. 2018. Vocalic, Lexical and Prosodic Cues for the INTERSPEECH 2018 Self-Assessed Affect Challenge. In *Interspeech 2018, 19th Annual Conference of the International Speech Communication Association, Hyderabad, India, 2-6 September 2018*, B. Yegnanarayana (Ed.). ISCA, 541–545. https://doi.org/10.21437/Interspeech.2018-1331

[35] Thi-Thu Nguyen and Seokhoon Yoon. 2019. A novel approach to short-term stock price movement prediction using transfer learning. *Applied Sciences* 9, 22 (2019), 4745.

[36] Mahla Nikou, Gholamreza Mansourfar, and Jamshid Bagherzadeh. 2019. Stock price prediction using DEEP learning algorithm and its comparison with machine learning algorithms. *Intelligent Systems in Accounting, Finance and Management* 26, 4 (2019), 164–174.

[37] Anne Opschoor, Dick van Dijk, and Michel van der Wel. 2014. Predicting volatility and correlations with financial conditions indexes. *Journal of Empirical Finance* 29 (2014), 435–447.

[38] Yaohao Peng, Pedro Henrique Melo Albuquerque, Jader Martins Camboim de Sá, Ana Julia Akaishi Padula, and Mariana Rosa Montenegro. 2018. The best of two worlds: Forecasting high frequency volatility for cryptocurrencies and traditional currencies with Support Vector Regression. *Expert Systems with Applications* 97 (2018), 177–192.

[39] Verónica Pérez-Rosas, Mohamed Abouelenien, Rada Mihalcea, Yao Xiao, CJ Linton, and Mihai Burzo. 2015. Verbal and nonverbal clues for real-life deception detection. In *Proceedings of the 2015 conference on empirical methods in natural language processing*. 2336–2346.

[40] Soujanya Poria, Erik Cambria, Devamanyu Hazarika, Navonil Majumder, Amir Zadeh, and Louis-Philippe Morency. 2017. Context-dependent sentiment analysis in user-generated videos. In *Proceedings of the 55th annual meeting of the association for computational linguistics (volume 1: Long papers)*. 873–883.

[41] 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*. Association for Computational Linguistics, Florence, Italy, 390–401. https://doi.org/10.18653/v1/P19-1038

[42] Harish Gandhi Ramachandran and Dan DeRose Jr. 2018. A text analysis of federal reserve meeting minutes. *arXiv preprint arXiv:1805.07851* (2018).

[43] Carlo Rosa. 2013. The financial market effect of FOMC minutes. *Available at SSRN 2378398* (2013).

[44] Francesco Rundo, Francesca Trenta, Agatino Luigi di Stallo, and Sebastiano Battiato. 2019. Machine learning for quantitative finance applications: A survey. *Applied Sciences* 9, 24 (2019), 5574.

[45] Ronnie Sadka. 2006. Momentum and post-earnings-announcement drift anomalies: The role of liquidity risk. *Journal of Financial Economics* 80, 2 (2006), 309–349.

[46] Hiroaki Sakoe and Seibi Chiba. 1978. Dynamic programming algorithm optimization for spoken word recognition. *IEEE transactions on acoustics, speech, and signal processing* 26, 1 (1978), 43–49.

[47] Ramit Sawhney, Mihir Goyal, Prakhar Goel, Puneet Mathur, and Rajiv Shah. 2021. Multimodal Multi-Speaker Merger & Acquisition Financial Modeling: A New Task, Dataset, and Neural Baselines. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint*

Conference on Natural Language Processing (Volume 1: Long Papers). 6751–6762.

[48] Ramit Sawhney, Puneet Mathur, Ayush Mangal, Piyush Khanna, Rajiv Ratn Shah, and Roger Zimmermann. 2020. Multimodal multi-task financial risk forecasting. In *Proceedings of the 28th ACM international conference on multimedia*. 456–465.

[49] Ramit Sawhney, Arnav Wadhwa, Shivam Agarwal, and Rajiv Shah. 2021. FAST: Financial News and Tweet Based Time Aware Network for Stock Trading. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. 2164–2175.

[50] Umut Mehmet Sen, Veronica Perez-Rosas, Berrin Yanikoglu, Mohamed Abouelenien, Mihai Burzo, and Rada Mihalcea. 2020. Multimodal deception detection using real-life trial data. *IEEE Transactions on Affective Computing* (2020).

[51] Adam Hale Shapiro and Daniel Wilson. 2021. Taking the fed at its word: A new approach to estimating central bank objectives using text analysis. Federal Reserve Bank of San Francisco.

[52] Mohammad Soleymani, Maja Pantic, and Thierry Pun. 2011. Multimodal emotion recognition in response to videos. *IEEE transactions on affective computing* 3, 2 (2011), 211–223.

[53] Raul Cruz Tadle. 2022. FOMC minutes sentiments and their impact on financial markets. *Journal of Economics and Business* 118 (2022), 106021.

[54] Yao-Hung Hubert Tsai, Shaojie Bai, Paul Pu Liang, J Zico Kolter, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2019. Multimodal transformer for unaligned multimodal language sequences. In *Proceedings of the conference. Association for Computational Linguistics. Meeting*, Vol. 2019. NIH Public Access, 6558.

[55] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is All you Need. In *Advances in Neural Information Processing Systems*, I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (Eds.), Vol. 30. Curran Associates, Inc. https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf

[56] Steven Walczak. 2001. An empirical analysis of data requirements for financial forecasting with neural networks. *Journal of management information systems* 17, 4 (2001), 203–222.

[57] Jerry Weiss. 2011. Ekman, P.(2009) Telling Lies: Clues to Deceit in the Marketplace, Politics, and Marriage. New York: Norton.

[58] 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)*. Association for Computational Linguistics, Melbourne, Australia, 1970–1979. https://doi.org/10.18653/v1/P18-1183

[59] 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)*. 1970–1979.

[60] Linyi Yang, Tin Lok James Ng, Barry Smyth, and Riuhai Dong. 2020. *HTML: Hierarchical Transformer-Based Multi-Task Learning for Volatility Prediction*. Association for Computing Machinery, New York, NY, USA, 441–451. https://doi.org/10.1145/3366423.3380128

[61] Li Yao, Atousa Torabi, Kyunghyun Cho, Nicolas Ballas, Christopher Pal, Hugo Larochelle, and Aaron Courville. 2015. Describing videos by exploiting temporal structure. In *Proceedings of the IEEE international conference on computer vision*. 4507–4515.

[62] ShuiLing Yu and Zhe Li. 2018. Forecasting stock price index volatility with LSTM deep neural network. In *Recent developments in data science and business analytics*. Springer, 265–272.

[63] Amir Zadeh, Chengfeng Mao, Kelly Shi, Yiwei Zhang, Paul Pu Liang, Soujanya Poria, and Louis-Philippe Morency. 2019. Factorized multimodal transformer for multimodal sequential learning. *arXiv preprint arXiv:1911.09826* (2019).