Modal-adaptive Knowledge-enhanced Graph-based Financial Prediction from Monetary Policy Conference Calls with LLM

Kun Ouyang[1], **Yi Liu**[1], **Shicheng Li**[1], **Ruihan Bao**[2], **Keiko Harimoto**[2], **Xu Sun**[1]
[1]National Key Laboratory for Multimedia Information Processing,
School of Computer Science, Peking University
[2]Mizuho Securities Co., Ltd.
kunouyang10@gmail.com, {imliuyi, lisc99}@pku.edu.cn,
{ruihan.bao, keiko.harimoto}@mizuho-sc.com, xusun@pku.edu.cn

Abstract

Financial prediction from Monetary Policy Conference (MPC) calls is a new yet challenging task, which targets at predicting the price movement and volatility for specific financial assets by analyzing multimodal information including text, video, and audio. Although the existing work has achieved great success using cross-modal transformer blocks, it overlooks the potential external financial knowledge, the varying contributions of different modalities to financial prediction, as well as the innate relations among different financial assets. To tackle these limitations, we propose a novel **M**odal-**A**daptive k**N**owledge-enh**A**nced **G**raph-bas**E**d financial p**R**ediction scheme, named MANAGER. Specifically, MANAGER resorts to FinDKG to obtain the external related knowledge for the input text. Meanwhile, MANAGER adopts BEiT-3 and Hidden-unit BERT (HuBERT) to extract the video and audio features, respectively. Thereafter, MANAGER introduces a novel knowledge-enhanced cross-modal graph that fully characterizes the semantic relations among text, external knowledge, video and audio, to adaptively utilize the information in different modalities, with ChatGLM2 as the backbone. Extensive experiments on a publicly available dataset Monopoly verify the superiority of our model over cutting-edge methods.

Keywords: Financial Prediction, LLM, Multimodal Learning

1. Introduction

Forecasting the fluctuation of prices for a financial asset over a specific period is a crucial task in financial analysis, essential for both investors and policymakers (Lewellen, 2003). Accurate prediction results can assist investors in making informed decisions regarding investment returns, while policymakers can implement prudent monetary measures to uphold a robust economy (Cai et al., 2021; Shapiro and Wilson, 2019). In early work, researchers made efforts to solve financial prediction for textual financial data, such as BloombergGPT (Wu et al., 2023b) and FinGPT (Wang et al., 2023b).

Despite their promising performance, the above models can only solve text-based financial tasks. With unprecedented advances in multimodal learning, investors now have access to a vast amount of unstructured data for financial prediction (Jiang, 2020). Moreover, the non-verbal information involved in the visual and acoustical modalities (e.g., vocal tone and facial expressions) can be indicative and correlated with trading activities in the financial market. One such abundant source of multimodal information is the Monetary Policy Conference (MPC's) call. Previous work (Boukus and Rosenberg, 2006) has underscored the influence of MPC calls on financial stock markets. Therefore, Mathur et al. (2022) curated a public Monetary Policy Call Dataset named Monopoly and proposed to predict the price movement and volatility for six principal financial assets (i.e., Stock In-

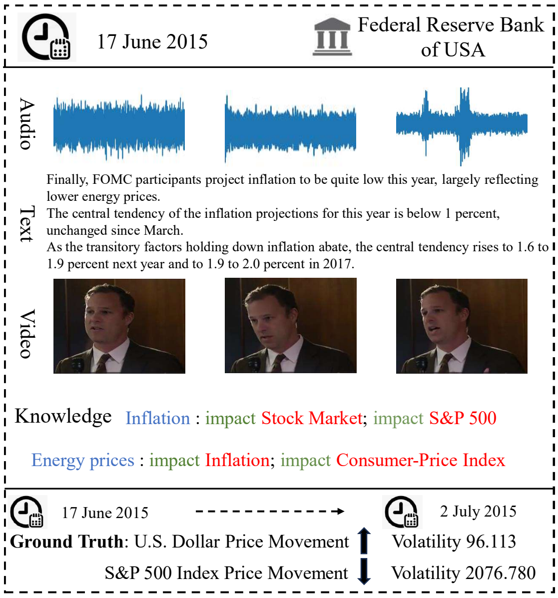

Figure 1: An example of the financial prediction from MPC calls. We also present the external knowledge inferred by FinDKG for the given text. Notably, the words in blue are the anchor entities while those in green are the relations and those in red are the related entities.

Proceedings of the Joint Workshop of the 7th FinNLP, the 5th KDF, and the 4th ECONLP, pages 59–69
20 May, 2024.

dex (Small), Stock Index (Large), Gold Price, Currency Exchange Rate, Long-term bond yield (10-years), Short-term bond yield (3-months)) based on MPC calls. The authors adopted cross-modal transformer blocks and modality-specific attention fusion to conduct price movement and volatility prediction. Although this pioneering study has achieved promising performance, it still suffers from three key limitations.

1) **Overlook the potential external knowledge in the financial domain.** The pioneering study fails to utilize the related knowledge contained in the external public knowledge base in the financial domain. As shown in Figure 1, the related knowledge obtained from FinDKG (Li, 2023) can strengthen the context comprehension (e.g., "impact S&P 500") and promote the financial prediction.

2) **Overlook the varying contributions of different modalities to financial prediction.** The existing work equally feeds the multimodal features (i.e., text, video and audio) into the model, and treats them as the equal source of information to conduct multimodal information fusion with the same weights. In fact, the content of given text is the prime cue for the financial prediction, while the non-verbal cues such as facial expressions and vocal tone involved in the video and audio play a minor role in comprehending the context. How to adaptively utilize the information residing in the multiple modalities merits our attention.

3) **Overlook the innate relations among different financial assets.** The former method predicts the price movement and volatility of six financial assets independently, ignoring the potential relationships among different financial assets. Actually, the price changes of a financial asset may provide useful information to predict price trend of the other financial assets.

To tackle these limitations, we propose a novel **M**odal-**A**daptive k**N**owledge-enh**A**nced **G**raph-bas**E**d financial p**R**ediction scheme, MANAGER for short. In detail, MANAGER consists of four components: external financial knowledge acquisition, video-audio feature extraction, knowledge-enhanced modal-adaptive context comprehension and task-specific instruction tuning for financial prediction, as shown in Figure 2. In the first module, we focus on acquiring the external related knowledge for the given text, where a large-scale financial knowledge base FinDKG (Li, 2023) is used. In the second module, we utilize BEiT-3 (Wang et al., 2022) and Hidden-Unit BERT (HuBERT) (Hsu et al., 2021) to extract the video and audio representations, respectively. In the third module, we construct the knowledge-enhanced cross-modal graph to aggregate the given text, input video, audio and inferred external knowledge through two types of relations (i.e., intra-modal and inter-modal semantic relations). We then employ the commonly used graph neural networks (GCNs) (Kipf and Welling, 2017), which have shown great performance in NLP tasks (Jing et al., 2023; Ouyang et al., 2024), to adaptively utilize the different modalities for cross-modal context comprehension. In the last module, considering that up-to-date Large Language Models (LLMs) have shown promising performance in multimodal context learning (Zhang and Li, 2023; Wu et al., 2023a), the potential of LLMs in the multimodal financial prediction task is increasingly evident. Therefore, we adopt ChatGLM2 (Du et al., 2022) as the backbone and feed the cross-modal representation into ChatGLM2 with a task-specific instruction devised for the certain financial asset to predict the price movement or volatility, respectively. Unlike previous work, we do not conduct prediction for different financial assets independently, but utilize ChatGLM2 to capture the innate relation among the financial assets. Finally, we conduct extensive experiments on a publicly available Monopoly dataset, on which our method outperforms the best baseline across all the metrics for both price movement and volatility prediction. Our contributions can be concluded as follows.

- We propose a novel modal-adaptive knowledge-enhanced graph-based financial prediction scheme, where the text, external knowledge, video and audio are aggregated for cross-modal context comprehension.
- As far as we know, we are the first to introduce an up-to-date LLM named ChatGLM2 to solve the financial prediction task for Monetary Policy Conference (MPC) calls, which contain multiple modalities (i.e., text, video and audio).
- The results of extensive experiments on the Monopoly dataset demonstrate the superiority of our MANAGER over other cutting-edge methods, and prove the effectiveness of each component of MANAGER. As a byproduct, we release our code and parameters[1] to facilitate the research community.

2. Related Work

2.1. Large Language Models (LLMs) in Finance

In early work about the application of LLMs in Finance, researchers resorted to BERT (Devlin et al., 2019) to conduct financial tasks, such as FinBert (Liu et al., 2020), which is dedicated for financial sentiment analysis with under one billion

[1] https://github.com/OuyangKun10/MANAGER.

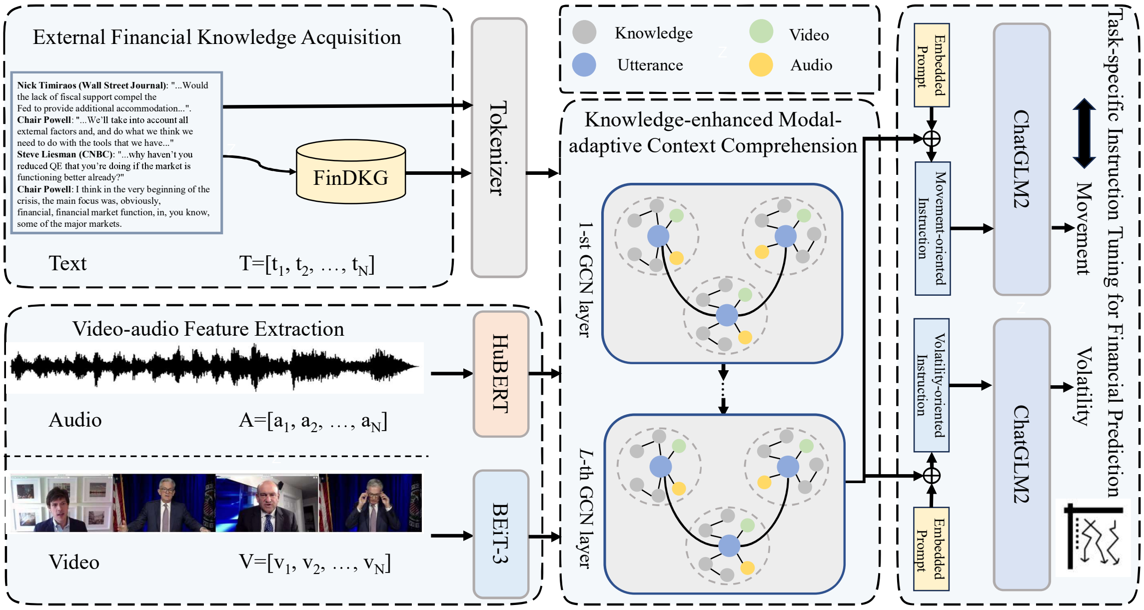

Figure 2: The architecture of MANAGER, which consists of four key components including External Financial Knowledge Acquisition, Video-audio Feature Extraction, Knowledge-enhanced Modal-adaptive Context Comprehension and Task-specific Instruction Tuning for Financial Prediction.

parameters, fine-tuned on a rich financial corpus to excel in finance-specific tasks. Although it achieves promising performance, it falls short of comprehending the long and complex financial text. In recent years, there has been a surge in research dedicated to integrating financial datasets with GPT-based models (Brown et al., 2020), aimed at enhancing Natural Language Processing (NLP) applications. For example, BloombergGPT (Wu et al., 2023b) is a closed-source model, trained extensively on diverse financial datasets, thereby encapsulating a broad spectrum of the financial domain. FinGPT (Wang et al., 2023b) is an open-source LLM, fine-tuned from a general LLM using low-rank adaptation method (Hu et al., 2021), fostering accessibility for the broader community.

2.2. Multimodal Financial Prediction

Existing work in the financial realm utilize vocal and textual cues from earnings conference calls (Qin and Yang, 2019; Sawhney et al., 2020), and mergers and acquisitions calls (Sawhney et al., 2021) for stock volatility prediction. Multimodal architectures that use these cues for financial predictions have seen significant improvements in their performances (Sawhney et al., 2020; Yang et al., 2020). However, the vision modality, which may offer important cues that correlate with the performance of financial markets (Cao, 2021) remains underexplored. Therefore, Mathur et al. (2022) first introduced video modality in the financial prediction task and released a dataset named Monopoly. They adopted cross-modal transformer blocks and modality-specific attention fusion to forecast the financial risk and price movement. Despite its promising performance on financial prediction, it overlooks the potential external knowledge, the varying contributions of different modalities, the innate relations among different financial assets, which are the major concerns of our model.

3. Task Formulation

Suppose we have a training dataset $\mathcal{D}$ composed of N_d samples, i.e., $\mathcal{D} = \{d_1, d_2, \cdots, d_{N_d}\}$. Each sample $d_i = \{T_i, V_i, A_i, Y_i\}$, where $T_i = \{u_1^i, u_2^i, \cdots u_N^i\}$ denotes the input text containing N utterances, $V_i = \{v_1^i, v_2^i, \cdots v_N^i\}$ and $A_i = \{a_1^i, a_2^i, \cdots a_N^i\}$ are the set of the input video and audio clips, respectively. Each utterance u_i contains N_{u_i} tokens. i.e., $u_i = \{t_1^i, t_2^i, \cdots, t_{N_{u_i}}^i\}$ And $Y_i^\tau = \{p_i^\tau, o_i^\tau\}$ denotes the target labels over a period of τ days, where p_i^τ is the price movement and o_i^τ is the volatility, respectively. Our target is to learn a multimodal financial prediction model $\mathcal{F}$ that is able to predict the price movement and volatility for six principal financial assets (i.e., Stock Index (Small), Stock Index (Large), Gold Price, Currency Exchange Rate, Long-term bond yield (10-years), Short-term bond yield (3-months)), based on the given multimodal input as follows,

$$\hat{Y}_i = \mathcal{F}(T_i, V_i, A_i|\Theta) \tag{1}$$

where Θ is a set of learnable parameters of the model $\mathcal{F}$. $\hat{Y}_i = \{\hat{p}_i^\tau, \hat{o}_i^\tau\}$ is the labels (i.e., price movement and volatility) predicted by $\mathcal{F}$. For simplicity, we omit the subscript i that indexes the training samples.

4. Method

In this section, we detail the four components of our proposed MANAGER, as shown in Figure 2.

4.1. External Financial Knowledge Acquisition

As aforementioned, the external financial knowledge inferred by the input text can assist the financial prediction since it can introduce corresponding financial entities as well as the relations, and provide some external factors to analyze the financial environment, leading to more informed predictions. Specifically, we resort to FinDKG (Li, 2023), which provides dynamic knowledge graph data in the financial domain, as the source of external knowledge. Notably, FinDKG changes dynamically over time. In detail, it contains $13,645$ financial entities and 15 types of relations. Given the input text, we adopt the period-specific FinDKG[2] that only contains the knowledge before the date of the input text, to prevent our model from obtaining the information beyond the date. The ration is that the information beyond the date can influence the prediction.

To acquire the related external knowledge for the given text, i.e., $T = \{u_1, u_2, \cdots u_N\}$, we first identify all the entities in FinDKG that exist in the input text. Let $\{e_1, \cdots, e_{N_e}\}$ be the set of identified entities, where N_e is the total number of the identified entities. We then use these identified entities as the anchors to obtain the related entities and corresponding relations as the external knowledge for the input text. Notably, for each anchor entity e, we retrieve all its one-hop neighboring entities, as well as the corresponding relations that are treated as the edges, from FinDKG and deem them as the external knowledge for e. Mathematically, let $\mathcal{N}(e) = \mathcal{N}^1(e) \cup \mathcal{N}^2(r)$ be the set of external knowledge (i.e., $\mathcal{N}^1(e)$ is the set of neighboring entities and $\mathcal{N}^2(r)$ is the set of corresponding relations between each neighboring entity and the anchor entity, respectively.) of the entity e in FinDKG. Then the related external knowledge for the input text can be represented as $\{\mathcal{N}^1_{e_1}, \mathcal{N}^1_{e_2}, \cdots, \mathcal{N}^1_{e_{N_e}}\} \cup \{\mathcal{N}^2_{e_1}, \mathcal{N}^2_{e_2}, \cdots, \mathcal{N}^2_{e_{N_e}}\}$. N_e is the number of the neighboring entities as well as the number of the relations.

[2] `https://xiaohui-victor-li.github.io/FinDKG/`.

4.2. Video-audio Feature Extraction

To obtain the global feature of video and audio clips, we choose BEiT-3 (Wang et al., 2022) and Hidden-Unit BERT (HuBERT) (Hsu et al., 2021) as the visual and acoustical encoder, respectively.

Video Encoding, to encode the video clips, we resort to BEiT-3, which is an advanced general-purpose multimodal foundation model pre-training for all vision and vision-language tasks and shows great performance in visual modality encoding (Wang et al., 2022). Specifically, we embed each frame v_j^k in the video clip 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 feature $x_V^j \in \mathbb{R}^D$, where D is the feature dimension. Mathematically, we have

$$x_V^j = \frac{1}{N_f} \sum_{k=1}^{N_f} \text{BEiT-3}(v_j^k), \forall j \in [1, N], \quad (2)$$

where N_f is the number of frames in the clips v_j. And we represent the sequence of video features as $X_V = [x_v^1, x_v^2, \cdots, x_v^N]$.

Audio Encoding, we extract the feature of the audio clips via the self-supervised speech representation model named HuBERT, which has shown significant power for extracting audio features for speech language understanding tasks (Yoon et al., 2022). We embed each audio utterance a_j^k in the audio clip a_j as the arithmetic mean of the representation derived by HuBERT and obtain the encoded acoustical feature $x_A^j \in \mathbb{R}^D$. Formally,

$$x_A^j = \text{HuBERT}(a_j^k), \forall j \in [1, N], \quad (3)$$

we represent the sequence of audio features as $X_A = [x_a^1, x_a^2, \cdots, x_a^N]$.

4.3. Knowledge-enhanced Modal-adaptive Context Comprehension

In this module, we aim to enhance the cross-modal context comprehension utilizing the inferred external knowledge in the financial domain. In fact, there are rich relations (i.e., intra-modal semantic relation and inter-modal semantic relation) existing in the multiple input including the text, video, audio and external knowledge. Therefore, to adaptively utilize different modalities via these semantic relations for boosting cross-modal context comprehension, we resort to the widely used graph neural networks (GCNs). Specifically, we first build a knowledge-enhanced cross-modal graph $\mathcal{G}$.

4.3.1. Nodes Initialization

In particular, the nodes in the knowledge-enhanced cross-modal graph $\mathcal{G}$ come from four kinds of sources, the given text T, input video clips V, input audio clips A and inferred external knowledge $\mathcal{N}(e)$. We define all the nodes as $\{n_1, \cdots, n_{N_c}\} = \{T, \mathcal{N}(e), V, A\}$, where N_c is the total number of nodes. To initialize the nodes, we feed the textual input $\{T, \mathcal{N}(e)\}$ into the encoder of ChatGLM2 (Du et al., 2022) to extract their features. Specifically, we first concatenate them into a sequence of tokens, denoted as $X_T = \{T, \mathcal{N}(e)\}$, and then feed X into the encoder $\mathcal{E}$ as follows,

$$\mathbf{H} = \mathcal{E}(X_T), \tag{4}$$

where $\mathbf{H} = [\mathbf{h}_1, \cdots, \mathbf{h}_{N_t+2\times N_e}] \in \mathbb{R}^{(N_t+2\times N_e)\times D}$ is the encoded representation matrix, N_t is the tokens number of the whole utterances and each column of which corresponds to a token. Accordingly, nodes of the textual part (utterances and external knowledge) in the knowledge-enhanced cross-modal graph $\mathcal{G}$ can be initialized by $\mathbf{H}$, where the j-th token node is initialized with $\mathbf{h}_j$. In addition, the other nodes are initialized by the extracted video feature sequence X_V and the extracted audio feature sequence X_A, respectively.

4.3.2. Semantic Relation Construction

To enhance the cross-modal context comprehension with related external knowledge, we consider two kinds of semantic relations: intra-modal semantic relation and inter-modal semantic relation. The former captures the basic information flow of the multiple modalities input, also incorporates the related external knowledge into the text. The latter enables the injection of non-textual information from video and audio into the context and achieves cross-modal information fusion.

Intra-modal Semantic Relation. To capture the information flow in the specific modality, we design three types of intra-modal semantic edges. a) *Token-token edges*. We introduce an edge between each pair of adjacent nodes in given text to represent the neighboring relations among the tokens of text. b) *Token-knowledge edge*. We connect the tokens that act as an anchor entity in the aforementioned external knowledge retrieval process, relation token and the related entity token sequentially. c) *Video-video edge* and d) *Audio-audio edge*. We link each pair of adjacent video nodes and connect each pair of adjacent audio nodes to represent the adjacency relations of the video and audio modalities, respectively. The above edges are characteristics of the information flow, and weighted by 1. Formally, we introduce the corresponding adjacency matrix $\mathbf{A}^1$ for representing these edges as follows,

$$\mathbf{A}^1_{i,j} = \begin{cases} 1, & if\ D_1(n_i, n_j), \\ 0, & otherwise, \end{cases} \tag{5}$$

where N_c denotes the total number of nodes in $\mathcal{G}$ and $i, j \in [1, N_c]$. $D_1(n_i, n_j)$ denotes that the nodes n_i and n_j have the certain above defined intra-modal semantic relation.

Inter-modal Semantic Relation. To comprehensively utilize the multiple modalities to promote cross-modal context comprehension, we devise two types of inter-modal semantic edges. a) *Token-video edges*. For each video node, we connect it to each token in the corresponding utterance. The ration is to inject the visual information (e.g., facial expressions and hand gestures) that can help semantics understanding and hence improve financial analysis, into the context. b) *Token-audio edges*. For each audio node, we link it with each token in the corresponding utterance. In this way, we can incorporate the acoustic information (e.g., vocal tone) that is also useful for context comprehension, into the context. The weight of all the above edges is set to 1. Accordingly, the adjacency matrix $\mathbf{A}^2 \in \mathbb{R}^{N_c \times N_c}$ for capturing the above inter-modal semantic relations can be constructed as follows,

$$\mathbf{A}^2_{i,j} = \begin{cases} 1, & if\ D_2(n_i, n_j), \\ 0, & otherwise, \end{cases} \tag{6}$$

where $D_2(n_i, n_j)$ indicates that nodes n_i and n_j have certain above inter-modal semantic semantic relation, $i \in [1, N_t]$ and $j \in [N_t + 2 \times N_e + 1, N_c]$.

Ultimately, by combing the adjacency matrices for intra-modal and inter-modal semantic relations, i.e., $\mathbf{A}^1$ and $\mathbf{A}^2$, we can derive the final adjacency matrix $\mathbf{A}$ for the knowledge-enhanced cross-modal graph.

4.3.3. Graph Convolution Network

Towards the final cross-modal context comprehension, we adopt L layers of GCN to extract the multimodal fusion feature of the cross-modal context. Then the node representations are iteratively updated as follows,

$$\mathbf{G}_l = ReLU(\tilde{\mathbf{A}}\mathbf{G}_{l-1}\mathbf{W}_l), l \in [1, L], \tag{7}$$

where $\tilde{\mathbf{A}} = (\mathbf{D})^{-\frac{1}{2}}\mathbf{A}(\mathbf{D})^{-\frac{1}{2}}$ is the normalized symmetric adjacency matrix, and $\mathbf{D}$ is the degree matrix of the adjacency matrix $\mathbf{A}$. In addition, $\mathbf{W}_l \in \mathbb{R}^{D\times D}$ is a trainable parameter of the l-th GCN layer. $\mathbf{G}_l$ are the node representations obtained by the l-th GCN layer, where $\mathbf{G}_0 = \mathbf{H}$ is the initial node representation.

4.4. Task-specific Instruction Tuning for Financial Prediction

The final nodes representation $\mathbf{G}_L$ obtained by the L-th layer GCNs absorb rich semantic information from their correlated nodes and can be used as the

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
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	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	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	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	0.435	<u>0.653</u>	0.302	0.605	0.392	<u>0.631</u>	0.329	0.601	0.296	0.217	0.142	0.385
MDRM	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	0.490	0.645	0.458	0.541	0.431	0.504	0.557	0.482	0.484	0.531	0.298	<u>0.626</u>
MULT	0.491	0.630	0.536	0.629	0.443	0.625	0.572	0.612	0.499	0.547	<u>0.473</u>	0.521
MPCNet	<u>0.501</u>	0.590	<u>0.565</u>	<u>0.638</u>	<u>0.460</u>	0.590	<u>0.559</u>	<u>0.620</u>	<u>0.520</u>	<u>0.570</u>	0.329	0.450
MANAGER	**0.548***	**0.694***	**0.610***	**0.659***	**0.517***	**0.652***	**0.589***	**0.646***	**0.564***	**0.608***	**0.511***	**0.681***

(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
HistPrice	0.360	0.390	0.350	0.400	0.310	0.290	0.220	0.390	0.220	0.160	0.340	0.330
P-SVM	0.390	0.420	0.370	0.380	0.340	0.310	0.330	0.330	0.370	0.220	0.310	0.390
P-LSTM	0.365	0.352	0.371	0.346	0.320	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	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	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	0.434	0.383	0.214	0.317	0.287	0.242	0.314	0.149	0.346	0.198	<u>0.478</u>	0.505
HTML	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	0.329	0.590	<u>0.454</u>	0.533	<u>0.534</u>	<u>0.364</u>	0.485	0.400	0.428	0.171	0.466	0.493
MPCNet	<u>0.444</u>	<u>0.668</u>	0.413	<u>0.637</u>	0.386	0.327	<u>0.560</u>	<u>0.625</u>	<u>0.493</u>	<u>0.556</u>	0.374	<u>0.537</u>
MANAGER	**0.486***	**0.696***	**0.507***	**0.672***	**0.612***	**0.391***	**0.587***	**0.649***	**0.521***	**0.583***	**0.519***	**0.574***

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

Table 1: Performance comparison with baselines for movement prediction in terms of F1 score τ-days after the call ($\tau \in \{1, 3, 7, 15\}$). The best results are in boldface, while the second best are underlined. $\star$ denotes that the p-value of the significant test between our result and the best baseline result is less than 0.01.

Model	Stock Index (Small)				Stock Index (Large)				Currency Exchange Rate			
	MSE_1	MSE_3	MSE_7	MSE_{15}	MSE_1	MSE_3	MSE_7	MSE_{15}	MSE_1	MSE_3	MSE_7	MSE_{15}
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	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	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	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	2.290	2.092	1.779	1.598	1.287	1.133	0.718	0.622	<u>0.975</u>	1.081	0.500	0.510
MDRM	<u>2.065</u>	2.511	1.748	1.597	1.281	1.578	0.683	0.612	1.183	1.627	0.769	0.512
HTML	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	2.073	2.179	1.768	1.605	1.288	1.133	<u>0.672</u>	0.742	1.022	1.018	0.549	0.497
MPCNet	2.233	<u>2.089</u>	<u>1.732</u>	<u>1.594</u>	<u>1.269</u>	<u>1.046</u>	0.806	<u>0.607</u>	1.176	<u>1.001</u>	<u>0.469</u>	<u>0.470</u>
MANAGER	**1.819***	**1.725***	**1.608***	**1.471***	**1.126***	**0.813***	**0.584***	**0.572***	**0.906***	**0.957***	**0.416***	**0.402***

(a) Stock Indices and Currency Exchange Rate

Model	Gold				10-Year Bond Yield				3-Month Bond Yield			
	MSE_1	MSE_3	MSE_7	MSE_{15}	MSE_1	MSE_3	MSE_7	MSE_{15}	MSE_1	MSE_3	MSE_7	MSE_{15}
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	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	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	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	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	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	<u>1.277</u>	1.291	0.589	0.524	<u>1.685</u>	1.612	1.103	1.149	2.342	2.356	1.962	1.998
MULT	1.314	1.335	0.579	0.503	2.122	1.837	1.104	<u>1.037</u>	<u>1.174</u>	2.515	1.973	1.903
MPCNet	1.342	<u>1.275</u>	<u>0.562</u>	<u>0.477</u>	1.767	<u>1.602</u>	<u>0.979</u>	1.142	2.431	<u>2.319</u>	<u>1.948</u>	<u>1.901</u>
MANAGER	**1.106***	**1.144***	**0.527***	**0.419***	**1.452***	**1.574***	**0.825***	**0.917***	**1.049***	**2.076***	**1.804***	**1.276***

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

Table 2: Performance comparison with baselines for volatility prediction in terms of MSE τ-days after the call ($\tau \in \{1, 3, 7, 15\}$). The best results are in boldface, while the second best are underlined. $\star$ denotes that the p-value of the significant test between our result and the best baseline result is less than 0.01.

input for the following financial prediction. Considering that we need to solve a couple of tasks (i.e., prediction of the price movement and volatility), we resort to the advanced large language model ChatGLM2, which shows great performance in context comprehension (Du et al., 2022), and fine-tune it for each task independently. In addition, constructing proper instructions is pivotal for task-specific tuning, with each task being guided by a unique instruction prompt. Therefore, we adopt the instruction template (Wang et al., 2023a) structured as follows:

`Instruction`: $[prompt]$ Input: $[input]$ Answer: $[output]$

This template provides a standardized format, facilitating consistency across different tasks. Notably, we utilize the aforementioned final nodes representation $\mathbf{G}_L$ as input. Next, we design prompt for specific task.

Movement Prediction. In this task, we aim to predict the price movement for the financial assets. Therefore, the movement-oriented prompt is designed to guide ChatGLM2 to judge the price movement (e.g., "increase" or "decrease") of the given asset based on the multimodal input. The prompt template is "Please predict the price movement of O in τ days after the $date$ according to the input", where O is the to-be-predicted financial asset, $\tau \in \{1, 3, 7, 15\}$ and $date$ is formatted as `YYYY-MM-DD`.

Volatility Prediction. In this task, we aim to predict the volatility, a float number that measures the instability of an asset. Therefore, the volatility-oriented prompt is designed to guide ChatGLM2 to output the volatility of the given financial asset based on the multimodal input. Similar to the above prompt template, we just replace "price movement" with "volatility".

We then utilize encoder of ChatGLM2 to embed the prompt and concatenate it with the input. $[output]$ is the prediction result that is answered after we feed the instruction into ChatGLM2. And we can obtain the task-specific instruction (i.e., Movement-oriented instruction I_p and volatility-oriented instruction I_v). Finally, we feed I_p and I_v into ChatGLM2 independently to guide it to conduct the two financial prediction tasks. For optimizing our MANAGER, we adopt Binary Cross-Entropy (BCE) loss and Mean Squared Error (MSE) loss to train the output for price movement prediction and volatility prediction, respectively.

5. Experiment

5.1. Dataset

In this work, we conducted extensive experiments on Monopoly (Mathur et al., 2022) dataset for financial prediction. It is a collection of public monetary conference call videos along with their corresponding audio recordings and text transcripts

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
w/o-Text	0.479	0.590	0.513	0.544	0.437	0.556	0.436	0.549	0.483	0.510	0.433	0.609
w/o-Knowledge	0.530	0.677	0.592	0.631	0.485	0.630	0.574	0.628	0.547	0.591	0.497	0.670
w/o-Video	0.509	0.664	0.581	0.627	0.477	0.606	0.562	0.611	0.530	0.573	0.484	0.657
w/o-Audio	0.527	0.671	0.586	0.612	0.493	0.639	0.570	0.620	0.546	0.588	0.494	0.660
w/o-Graph	0.533	0.679	0.491	0.608	0.492	0.611	0.560	0.627	0.532	0.570	0.483	0.654
w/-FullGraph	0.499	0.582	0.548	0.601	0.429	0.627	0.451	0.595	0.510	0.521	0.448	0.639
MANAGER	**0.548**	**0.694**	**0.610**	**0.659**	**0.517**	**0.652**	**0.589**	**0.646**	**0.564**	**0.608**	**0.511**	**0.681**

(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
w/o-Text	0.369	0.571	0.448	0.525	0.498	0.328	0.467	0.583	0.460	0.516	0.428	0.502
w/o-Knowledge	0.470	0.659	0.488	0.656	0.593	0.370	0.562	0.627	0.504	0.576	0.498	0.557
w/o-Video	0.451	0.639	0.471	0.646	0.589	0.352	0.550	0.617	0.501	0.537	0.500	0.545
w/o-Audio	0.456	0.662	0.479	0.631	0.573	0.362	0.554	0.619	0.496	0.553	0.481	0.521
w/o-Graph	0.457	0.681	0.457	0.650	0.593	0.359	0.559	0.624	0.503	0.551	0.487	0.540
w/-FullGraph	0.397	0.592	0.461	0.620	0.519	0.317	0.512	0.607	0.495	0.532	0.413	0.518
MANAGER	**0.486**	**0.696**	**0.507**	**0.672**	**0.612**	**0.391**	**0.587**	**0.649**	**0.521**	**0.583**	**0.519**	**0.574**

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

Table 3: Ablation study results of our proposed MANAGER for movement prediction. The best results are highlighted in boldface.

Model	Stock Index (Small)				Stock Index (Large)				Currency Exchange Rate			
	MSE_1	MSE_3	MSE_7	MSE_{15}	MSE_1	MSE_3	MSE_7	MSE_{15}	MSE_1	MSE_3	MSE_7	MSE_{15}
w/o-Text	2.146	2.048	1.795	1.613	1.379	1.016	0.879	0.926	1.198	1.348	0.702	0.617
w/o-Knowledge	1.930	1.908	1.784	1.536	1.291	0.905	0.670	0.713	0.977	1.019	0.474	0.566
w/o-Video	2.101	1.997	1.891	1.741	1.324	1.089	0.851	0.854	1.122	1.141	0.668	0.609
w/o-Audio	1.959	1.893	1.748	1.609	1.237	1.003	0.776	0.790	1.035	1.124	0.603	0.593
w/o-Graph	1.941	1.929	1.754	1.712	1.331	1.004	0.750	0.755	1.136	1.180	0.603	0.449
w/-FullGraph	2.144	1.962	1.919	1.791	1.495	1.022	0.898	0.904	1.175	1.306	0.681	0.640
MANAGER	**1.819**	**1.725**	**1.608**	**1.471**	**1.126**	**0.813**	**0.584**	**0.572**	**0.906**	**0.957**	**0.416**	**0.402**

(a) Stock Indices and Currency Exchange Rate

Model	Gold				10-Year Bond Yield				3-Month Bond Yield			
	MSE_1	MSE_3	MSE_7	MSE_{15}	MSE_1	MSE_3	MSE_7	MSE_{15}	MSE_1	MSE_3	MSE_7	MSE_{15}
w/o-Text	1.451	1.400	0.796	0.794	1.812	1.910	1.110	1.173	1.400	2.403	2.090	1.498
w/o-Knowledge	1.222	1.215	0.593	0.427	1.594	1.708	0.886	1.092	1.237	2.211	1.900	1.467
w/o-Video	2.184	1.351	0.576	0.704	1.641	1.850	1.070	1.101	1.344	2.280	1.979	1.510
w/o-Audio	1.307	1.358	0.639	0.549	1.588	1.758	1.024	1.127	1.264	2.213	1.943	1.419
w/o-Graph	1.347	1.278	0.742	0.540	0.674	1.861	1.013	1.164	1.275	2.226	1.937	1.494
w/-FullGraph	1.365	1.367	0.795	0.680	1.816	1.834	1.073	1.168	1.234	2.373	2.075	1.463
MANAGER	**1.106**	**1.144**	**0.527**	**0.419**	**1.452**	**1.574**	**0.825**	**0.917**	**1.049**	**2.076**	**1.804**	**1.276**

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

Table 4: Ablation study results of our proposed MANAGER for volatility prediction. The best results are highlighted in boldface.

released by six international banks between 2009 and 2022. Overall, it consists of $24,180$ samples, and each sample includes the corresponding text, video and audio clips with the annotated price movement and volatility. We adopted the original dataset split setting, the ratio of data split for training/validation/testing sets is $7:1:2$.

5.2. Experimental Setup

We adopted ChatGLM2[3] as the backbone of our model. The total number of tokens in each sample, i.e., N_t is unified to 768. The feature dimension D is set to 768. We used AdamW (Loshchilov and Hutter, 2017) as the optimizer and set the learning rate of GCN layers to 1e-3. Following Mathur et al. (2022), we use a learning rate of 1e-4 for movement prediction and 1e-3 for volatility prediction, respectively. The batch size is set to 1 due to GPU limitation, and the maximum number of epochs for model training is set to 10. Following the previous work, we employed mean squared error (MSE) to evaluate the predicted volatility and used F1 score to measure the predicted price movement, respectively, for $\tau \in \{1, 3, 7, 15\}$.

5.3. Baseline methods

5.3.1. Text-only baselines

- **HistPrice** (Du and Budescu, 2007). It utilizes the ARIMA model to perform regression/classification.

[3] https://huggingface.co/THUDM/chatglm2-6b.

- **P-SVM** (Chatzis et al., 2018). This model applies Support Vector Regression (SVR) and Classifiers (SVC) for volatility and price movement prediction, respectively.
- **P-LSTM** (Yu and Li, 2018). It uses LSTM to extract forecast patterns from 30-day historical price time-series.

5.3.2. Multimodal baselines

- **MLP** (Tolstikhin et al., 2021). It is a simple multi-layer perceptron where multimodal features are aggregated across a time series and concatenated for prediction.
- **LSTM** (Poria et al., 2017). It feeds the multimodal time series to individual LSTMs and averages them before the final prediction.
- **MMIM** (Han et al., 2021). In this model, LSTMs are employed to encode the video and audio sequences, while BERT is utilized for text processing. Subsequently, the encoded features are fused for prediction.
- **MDRM** (Qin and Yang, 2019). It adopts BiLSTM layers to encode unimodal sequences, which are then fused together for prediction.
- **HTML** (Yang et al., 2020). HTML utilizes fused multimodal feature representations before passing through Transformer layers for final prediction.
- **MULT** (Tsai et al., 2019). It employs transformer encoders to align text, video, and audio sequences for prediction.

- **MPCNet** (Mathur et al., 2022). It adopts cross-modal transformer blocks and modality-specific attention fusion for prediction.

5.4. Experimental results

We reported the experiment results in Table 1 and Table 2. From the above tables, we have the following observations. 1) Our model **MANAGER** consistently exceeds all the baselines in terms of all the metrics for both price movement and volatility prediction, which thoroughly demonstrates the superiority of our model. 2) Overall, the second best model is always multimodal baseline which verifies that the video and audio modalities can provide useful information for the financial prediction. 3) Notably, multimodal models not always outperform text-only models. For example, **HistPrice**, **P-SVM** and **P-LSTM** exceed **MLP** in the movement prediction of Stock Index (Large). It imples that improper use of non-verbal information in video and audio may lead to worse performance.

6. Analyses

6.1. Ablation Study

We introduced the following variants to explore the contribution of each component.

- **w/o-Text**, **w/o-Knowledge**, **w/o-Video w/o-Audio** and **w/o-Graph**. To prove the effectiveness of the input text, inferred knowledge, video, audio and constructed knowledge-enhanced cross-modal graph, we eliminated the text, external financial knowledge, video, audio and graph in these variants, respectively.

- **w/-FullGraph**. To further investigate the semantic relations of our knowledge-enhanced cross-modal graph, we erased all the semantic relations and transformed the semantic graph to a full connection graph.

The ablation study results are shown in Table 3 and Table 4. From this table, we have the following observations. 1) w/o-Text performs terribly compared with MANAGER. This is reasonable since the caption is the main source for delivering information to predict the price movement or volatility. 2) MANAGER exceeds w/o-Knowledge. It proves that external knowledge in the financial domain can assist in comprehending the context. 3) MANAGER consistently outperforms w/o-Video and w/o-audio across different evaluation metrics. It demonstrates the non-verbal information residing in the video and audio can improve context comprehension and hence boost financial prediction. 4)w/o-Text performs worse than both w/o-Video and w/o-Audio. It implies that the given text contributes more to the financial prediction than video and audio. 5) MANAGER outperforms w/o-Graph, denoting that the graphs are essential to capture the semantic relations among text, knowledge, video and audio, which help comprehend the cross-modal context. And 6) w/-FullGraph performs worse than MANAGER, which verifies the utility of proposed semantic relations.

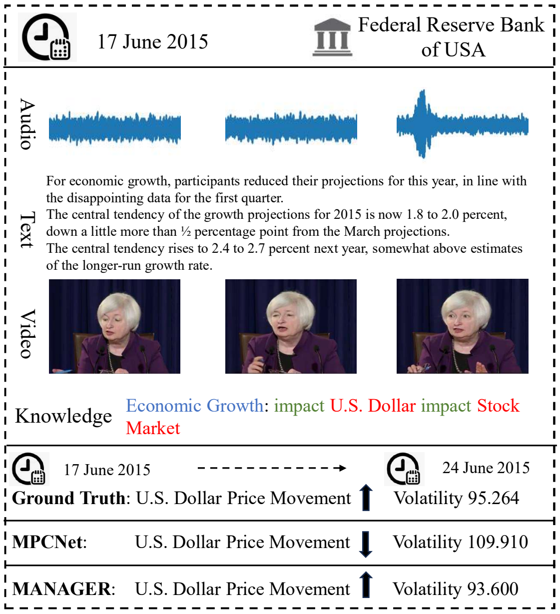

Figure 3: Comparison between the results predicted by MANAAGER and the best baseline MPCNet on one testing sample.

6.2. Case Study

To get an intuitive understanding of how our model works on financial prediction from MPC calls, we showed one testing sample in Figure 3 due to the limited space. For comparison, we also displayed the prediction results of the best baseline MPCNet.

As you can see, our MANAGER predicted the price movement of U.S. Dollar correctly, while MPCNet failed. In addition, the volatility 93.600 forecasted by MANAGER is closer to the ground truth 95.264 than 109.910 predicted by MPCNet. This may be attributed to the fact that the external knowledge (e.g., relation: "impact", entity: "U.S. Dollar" and "Stock Market") inferred by the entity "Economic Growth" may guide our model to pay attention to "Economic Growth" existed in the text, since it may provide some useful information for the price movement or volatility of U.S. Dollar. Overall, this case shows the benefit of incorporating external knowledge into the context of financial prediction from MPC calls.

7. Conclusion and Future Work

In this work, we propose a novel modal-adaptive knowledge-enhanced graph-based financial prediction scheme. Experimental results on a public dataset demonstrate the superiority of our model over existing cutting-edge methods, and validate the advantage of utilizing external knowledge in the financial domain, as well as the benefit of constructing the knowledge-enhanced cross-modal graph to characterize the intra-modal and inter-modal relations among the multiple input (i.e., text, external knowledge, video and audio). In the future, we plan to explore the Multimodal Large Language Models, such as VisualGLM, in financial prediction.

8. Acknowledgement

We thank all the anonymous reviewers for their valuable suggestions. This work is supported by a Research Grant from Mizuho Securities Co., Ltd. We sincerely thank Mizuho Securities for valuable domain expert suggestions. Ruihan Bao and Xu Sun are the corresponding authors.

9. Bibliographical References

Ellyn R. Boukus and Joshua V. Rosenberg. 2006. The information content of fomc minutes. *Monetary Economics*.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, T. J. Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeff Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *ArXiv*.

Yongbin 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.

Longbing Cao. 2021. Ai in finance: Challenges, techniques, and opportunities. *ACM Computing Surveys (CSUR)*.

Sotirios P. Chatzis, Vasilis Siakoulis, Anastasios Petropoulos, Evangelos Stavroulakis, and Nikos E. Vlachogiannakis. 2018. Forecasting stock market crisis events using deep and statistical machine learning techniques. *Expert Syst. Appl.*

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 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4171–4186. ACL.

Ning Du and David V. Budescu. 2007. Does past volatility affect investors' price forecasts and confidence judgements? *International Journal of Forecasting*.

Zhengxiao Du, Yujie Qian, Xiao Liu, Ming Ding, Jiezhong Qiu, Zhilin Yang, and Jie Tang. 2022. Glm: General language model pretraining with autoregressive blank infilling. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, pages 320–335.

Wei Han, Hui Chen, and Soujanya Poria. 2021. Improving multimodal fusion with hierarchical mutual information maximization for multimodal sentiment analysis. *ArXiv*.

Wei-Ning Hsu, Benjamin Bolte, Yao-Hung Hubert Tsai, Kushal Lakhotia, Ruslan Salakhutdinov, and Abdel rahman Mohamed. 2021. Hubert: Self-supervised speech representation learning by masked prediction of hidden units. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*.

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

Weiwei Jiang. 2020. Applications of deep learning in stock market prediction: recent progress. *Expert Syst. Appl.*

Liqiang Jing, Xuemeng Song, Kun Ouyang, Mengzhao Jia, and Liqiang Nie. 2023. Multi-source semantic graph-based multimodal sarcasm explanation generation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 11349–11361. ACL.

Thomas N. Kipf and Max Welling. 2017. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*. OpenReview.net.

Katharina Lewellen. 2003. Financing decisions when managers are risk averse. *MIT Sloan School of Management Working Paper Series*.

Xiaohui Victor Li. 2023. Findkg: Dynamic knowledge graph with large language models for global finance. pages 1–64. SSRN.

Vivian Liu and Lydia B. Chilton. 2021. Design guidelines for prompt engineering text-to-image generative models. *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*.

Zhuang Liu, Degen Huang, Kaiyu Huang, Zhuang Li, and Jun Zhao. 2020. Finbert: A pre-trained financial language representation model for financial text mining. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI-20*, pages 4513–4519. International Joint Conferences on Artificial Intelligence Organization.

Ilya Loshchilov and Frank Hutter. 2017. Fixing weight decay regularization in adam. *CoRR*, abs/1711.05101.

Puneet Mathur, Atula Tejaswi 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. *Proceedings of the 30th ACM International Conference on Multimedia*.

Kun Ouyang, Liqiang Jing, Xuemeng Song, Meng Liu, Yupeng Hu, and Liqiang Nie. 2024. Sentiment-enhanced graph-based sarcasm explanation in dialogue. *CoRR*, abs/2402.03658.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke E. Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Francis Christiano, Jan Leike, and Ryan J. Lowe. 2022. Training language models to follow instructions with human feedback. *ArXiv*.

Soujanya Poria, E. Cambria, Devamanyu Hazarika, Navonil Majumder, Amir Zadeh, and Louis-Philippe Morency. 2017. Context-dependent sentiment analysis in user-generated videos. In *Annual Meeting of the Association for Computational Linguistics*.

Yu Qin and Yi Yang. 2019. What you say and how you say it matters: Predicting stock volatility using verbal and vocal cues. In *Annual Meeting of the Association for Computational Linguistics*.

Ramit Sawhney, Mihir Goyal, Prakhar Goel, Puneet Mathur, and Rajiv Ratn Shah. 2021. Multimodal multi-speaker merger & acquisition financial modeling: A new task, dataset, and neural baselines. In *Annual Meeting of the Association for Computational Linguistics*.

Ramit Sawhney, Puneet Mathur, Ayush Mangal, Piyush Khanna, Rajiv Ratn Shah, and Roger Zimmermann. 2020. Multimodal multi-task financial risk forecasting. *Proceedings of the 28th ACM International Conference on Multimedia*.

Adam Hale Shapiro and Daniel J. Wilson. 2019. Taking the fed at its word: A new approach to estimating central bank objectives using text analysis. *Federal Reserve Bank of San Francisco, Working Paper Series*.

Ilya O. Tolstikhin, Neil Houlsby, Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Thomas Unterthiner, Jessica Yung, Daniel Keysers, Jakob Uszkoreit, Mario Lucic, and Alexey Dosovitskiy. 2021. Mlp-mixer: An all-mlp architecture for vision. In *Neural Information Processing Systems*.

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. *Proceedings of the conference. Association for Computational Linguistics. Meeting*.

Neng Wang, Hongyang Yang, and Chris Wang. 2023a. Fingpt: Instruction tuning benchmark for open-source large language models in financial datasets. *ArXiv*.

Neng Wang, Hongyang Yang, and Christina Dan Wang. 2023b. Fingpt: Instruction tuning benchmark for open-source large language models in financial datasets. *NeurIPS Workshop on Instruction Tuning and Instruction Following*.

Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, and Furu Wei. 2022. Image as a foreign language: Beit pretraining for all vision and vision-language tasks. *CoRR*, abs/2208.10442.

Jules White, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith, and Douglas C. Schmidt. 2023. A prompt pattern catalog to enhance prompt engineering with chatgpt. *ArXiv*.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Shaokun Zhang, Erkang Zhu, Beibin Li, Li Jiang, Xiaoyun Zhang, and Chi Wang. 2023a. Autogen: Enabling next-gen llm applications via multi-agent conversation framework. *ArXiv*.

Shijie Wu, Ozan Irsoy, Steven Lu, Vadim Dabravolski, Mark Dredze, Sebastian Gehrmann, Prabhanjan Kambadur, David Rosenberg, and Gideon Mann. 2023b. Bloomberggpt: A large language model for finance. *ArXiv*.

Linyi Yang, Tin Lok James Ng, Barry Smyth, and Ruihai Dong. 2020. Html: Hierarchical transformer-based multi-task learning for volatility prediction. *Proceedings of The Web Conference 2020*.

J. Yoon, Beom Jun Woo, and Nam Soo Kim. 2022. Hubert-ee: Early exiting hubert for efficient speech recognition. *ArXiv*.

Shui-Ling Yu and Zhe Li. 2018. Forecasting stock price index volatility with lstm deep neural network.

Yixuan Zhang and Haonan Li. 2023. Can large langauge model comprehend ancient chinese? a preliminary test on aclue. In *International Conference on Algebraic and Logic Programming*.