
FinMultiTime: A Four-Modal Bilingual Dataset for Financial Time-Series Analysis

Wenyan Xu^{1*} Dawei Xiang^{2*} Yue Liu³ Xiyu Wang³ Yanxiang Ma³
 Liang Zhang³ Chang Xu³ Jiaheng Zhang³

¹Central University of Finance and Economics ²University of Connecticut

³National University of Singapore, The University of Sydney, HEC Paris

{2022211032@email.cufe.edu.cn, ieb24002@uconn.edu, yliu@u.nus.edu}
 {xiyuwang.usyd@gmail.com, yama9404@uni.sydney.edu.au}
 {zhangl@hec.fr, c.xu@sydney.edu.au, jhzhang@nus.edu.sg}

Abstract

Pure time-series forecasting tasks typically focus exclusively on numerical features; however, real-world financial decision-making demands the comparison and analysis of heterogeneous sources of information. Recent advances in deep learning and large-scale language models (LLMs) have made significant strides in capturing sentiment and other qualitative signals, thereby enhancing the accuracy of financial time-series predictions. Despite these advances, most existing datasets consist solely of price series and news text, are confined to a single market, and remain limited in scale. In this paper, we introduce **FinMultiTime**, the first large-scale, multimodal financial time-series dataset. FinMultiTime temporally aligns four distinct modalities—financial news, structured financial tables, K-line technical charts, and stock price time series—across both the S&P 500 and HS 300 universes. Covering 5,105 stocks from 2009 to 2025 in the United States and China, the dataset totals 112.6 GB and provides minute-level, daily, and quarterly resolutions, thus capturing short-, medium-, and long-term market signals with high fidelity. Our experiments demonstrate that (1) scale and data quality markedly boost prediction accuracy; (2) multimodal fusion yields moderate gains in Transformer models; and (3) a fully reproducible pipeline enables seamless dataset updates. The data for this paper can be found at².

1 Introduction

Time-series regression models have long been the cornerstone of financial valuation and forecasting. Traditional statistical approaches [30, 2, 3] focus exclusively on numerical features and overlook open-domain knowledge from diverse modalities [7]. Intuitively, fusing information across these modalities yields richer, multidimensional representations that can outperform uni-modal models [11]. In equity investment decisions, for example, investors integrate historical price series with real-time multimodal data to guide buy, sell, or hold strategies: structured tables supply fundamental metrics, news sentiment reflects market mood, and technical charts quantify long-term trends via indicators such as moving averages [38, 39].

Moreover, according to the Efficient Market Hypothesis [25], prices absorb information with a lag, which provides a theoretical basis for exploiting multi-source signals not yet fully reflected in

*Equal contribution.

²https://huggingface.co/datasets/Wenyan0110/Multimodal-Dataset-Image_Text_Table_TimeSeries-for-Financial-Time-Series-Forecasting

Table 1: Comparison of existing multimodal financial time-series datasets.

Dateset Benchmarks	Domain	Language	Text	Time Series	Image	Table	Span	Finest Frequency
Time-MMD [20]	Multi-domain (Economics)	English	✓	✓	✗	✗	1989-2024	Monthly
CiK [31]		English	✓	✓	✗	✗	2024	Monthly
NewsForecast [29]	Multi-domain (Bitcoin)	English	✓	✓	✗	✗	2019-2021	Daily
TimeCAP [19]		English	✓	✓	✗	✗	2019-2023	Daily
TSQA [15]	Multi-domain (Finance)	English	✓	✓	✗	✗	—	—
FNSPID Nasdaq [8]		English, Russian	✓	✓	✗	✗	2009-2023	Minute-Level
ACL18 [35]		English	✓	✓	✗	✗	2014-2016	Minute-Level
CIKM18 [32]	Finance	English	✓	✓	✗	✗	2017	Minute-Level
DOW30 [6]		English	✓	✓	✗	✗	2020-2022	Daily
Emnlp24 findings [16]		English	✓	✓	✗	✓	2010-2020	Quarterly
SEP [14]		English	✓	✓	✗	✗	2020-2022	Minute-Level
FinBen [33]		English, Spanish	✓	✓	✗	✗	—	—
FinMultiTime (Ours)		English, Chinese	✓	✓	✓	✓	2009-2025	Minute-Level

stock prices to predict future movements. Consequently, robust and reliable predictive models must assimilate heterogeneous data to capture the full complexity of price dynamics [4, 9].

Recently, the natural language processing (NLP) models enable sentiment analysis of financial news, event extraction from disclosures, table parsing in earnings reports, and automated chart summarization [28, 1, 36, 5, 18]. Despite rapid advances in NLP models, existing multimodal datasets remain constrained. Most integrate only price and sentiment within a single market, risking information loss (Table 1); recent efforts [16] incorporate quarterly tables but suffer from limited temporal coverage and low update frequency. Such datasets are too small to train large models or to validate generalization across market regimes [22, 26, 21], and they exacerbate large language models’ propensity for hallucinations in rapidly evolving financial environments [10, 37].

To address these limitations, we introduce FinMultiTime, a bilingual, large-scale dataset. FinMultiTime temporally aligns data from year 2009 to 2025 by four modalities, including text, tables, images, and time series. Our dataset includes 4213 S&P 500 constituents and 892 HS 300 constituents. After rigorous cleaning and preprocessing, FinMultiTime comprises 112.6 GB of minute, daily and quarterly level data covering both U.S (Table 2), and Chinese markets. Real-time updates ensure the dataset reflects the latest market conditions, providing a comprehensive foundation for developing and validating multimodal forecasting models. Experimental results demonstrate that incorporating large-scale multimodal data significantly reduces prediction error and improves trend-direction accuracy, with high-quality sentiment and long-term trend information proving especially critical.

Table 2: Overview of Bilingual Financial Dataset Specifications for the HS300 (Chinese) and S&P 500 (English) Indices

Bilingual Dataset	Type	Size	Format	Stocks	Records	Frequency
HS300 (Chinese)	Image	2.43 GB	PNG	810	52,914	Semi-Annual
	Table	568 MB	JSON/JSONNL	810	2,430	Quarterly/Annual
	Time series	345 MB	CSV	810	810	Daily
	Text	652.53 MB	JSONNL	892	1,420,362	Minute-Level
	All	3.96 GB	—	—	1,476,516	—
SP500 (English)	Image	8.67 GB	PNG	4,213	195,347	Semi-Annual
	Table	84.04 GB	JSON/JSONNL	2,676	8,028	Quarterly/Annual
	Time series	1.83 GB	CSV	4,213	4,213	Daily
	Text	14.1 GB	JSONNL	4,694	3,351,852	Minute-Level
	All	108.64 GB	—	—	3,559,440	—

2 Constructing FinMultiTime

The construction of the FinMultiTime dataset begins with the systematic acquisition and processing of multi-source information. In this section, we detail the sources and procedures involved in assembling all modalities of FinMultiTime as shown in Figure 1.

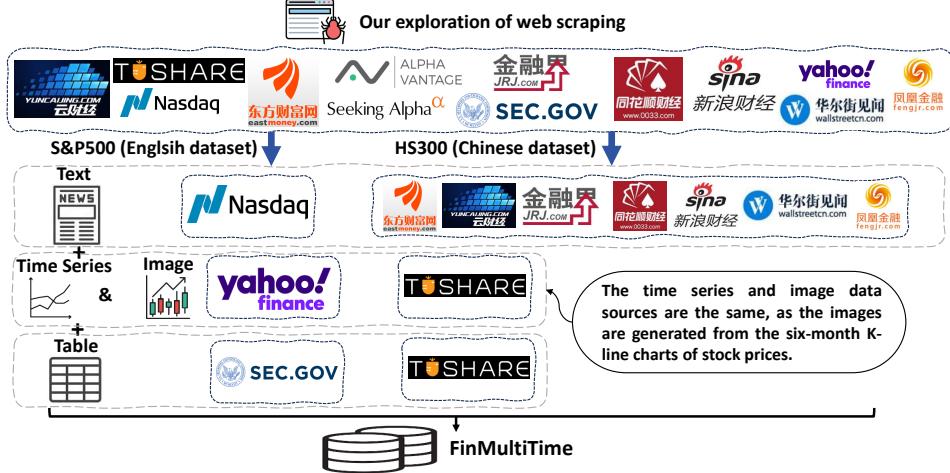


Figure 1: Data Collection Pipeline for the Bilingual Four-Modal FinMultiTime Dataset

Table 3: Comparison of Financial Tables for HS300 and S&P 500. The 10-Q is a quarterly financial report filed by publicly traded companies, while the 10-K is a comprehensive annual report. Both provide detailed information on a company’s financial position, operating performance, and cash flow at the end of the reporting period.

Table Type	HS300 (Chinese)			S&P500 (English)		
	Balance Sheet	Cash Flow Statement	Income Statement	Balance Sheet	Cash Flow Statement	Equity Statement
Format	JSONL	JSON	JSONL	JSON	JSONL	JSON
Field Count	147	31	92	28	80	33
10-Q nums	48,537	81,070	45,257	81,070	47,526	81,070
10-K nums	24,551	27,793	18,260	27,793	18,636	27,793
Time span	2001/12/31-2024/09/30			2000/01/03-2025/04/25		

2.1 Data Collection

We collect data from two of the major financial markets, as shown in Table 2. For **the U.S.** stock **numerical** data, we first retrieve daily OHLCV (*Open, High, Low, Close, and Volume*) data for S&P 500 constituent stocks via the Yahoo Finance API³. We then segment the data into semi-annual windows and generate candlestick **charts** using the mplfinance library. In these charts, a red candlestick body indicates that the closing price exceeded the opening price on a given day, while a green body indicates the opposite. The volume bars are color-coded to match the direction of price movement, offering an intuitive visual correlation between price trends and trading volume. For **news** sentiment data, due to strict usage restrictions imposed by various platforms (e.g., Investing.com, Seeking Alpha, and Alpha Vantage), we adopt the Nasdaq news scraping strategy from the FNSPID project[8] and implement several enhancements. These include improved handling of abnormal pages, refined auto-pagination logic, cookie popup filtering, and adaptation to different versions of ChromeDriver. The original scripts are upgraded into a robust, continuously running pipeline, substantially minimizing the risk of crawler interruption. The entire news scraping process is divided into two phases: the first leverages Selenium to collect news headlines and corresponding URLs for each stock; the second extracts full article content from these URLs, ultimately forming the text modality of the dataset. Structured financial **tables** are obtained primarily via the SEC Submissions and Company Facts APIs⁴. From 10-K and 10-Q filings of S&P 500 companies since 2000, we automatically extract key indicators from XBRL facts in balance sheets, cash flow statements, and statements of shareholders’ equity, while removing irrelevant fields such as announcement dates and filing types. For details on the retrieved tabular data, see Table 3.

For **the Chinese** market, daily **numerical** OHLCV data for HS 300 constituent stocks is retrieved through the Tushare API⁵ and used to generate technical candlestick **charts** consistent with the

³<https://finance.yahoo.com/>

⁴<https://www.sec.gov/search-filings/edgar-application-programming-interfaces>

⁵<https://tushare.pro/>

Table 4: Comparison of Two News Sources and Data Attributes

Source	Nasdaq News	Sina Finance	WallstreetCN	10jqka	Eastmoney	Yuncaijing	Fenghuang	Jinrongjie
Time Period	2009-04-08 to 2025-04-08				2020-03-31 to 2025-03-31			
Stock Symbol	Yes	No	No	No	No	No	No	No
Headline	Yes	No	Yes	Yes	Yes	Yes	No	Yes
URL	Yes	No	No	No	No	No	No	No
Text Type	Article				Flash News			
Filter Rate	–	18.12%	14.83%	22.51%	21.20%	53.39%	19.57%	24.35%
Summarization	No	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Language	English	Chinese	Chinese	Chinese	Chinese	Chinese	Chinese	Chinese

U.S. market. **News** sentiment data is also collected via the Tushare API, covering the period from March 31, 2020, to March 31, 2025. The dataset incorporates Chinese news from sources including Sina Finance, Wallstreetcn, iFinD, Eastmoney, YunCaijing, Ifeng News, and JRJ. Detailed bilingual news statistics are presented in Table 4. Structured financial **table** data for the HS 300 is acquired via Tushare API, including quarterly and annual balance sheets, income statements, and cash flow statements for the period spanning 2005 to 2024.

Data Ethics To ensure ethical compliance, we strictly adhere to the directives specified in robots.txt files during the U.S. news crawling process, collecting only publicly available content that does not require payment or subscription. Although Nasdaq does not offer an official API, web scraping is performed solely based on prior authorized research work, and all processed data is used exclusively within the FinMultiTime framework. For Chinese news, in order to avoid potential copyright and conflict-of-interest issues, we extract and utilize summaries of articles retrieved via the Tushare API, without publicly disclosing the full original content.

2.2 Data Mining and Preprocessing

To construct FinMultiTime, we extract and align four distinct data modalities—technical chart images, structured financial tables, normalized price series, and news text—across mostly constituent stocks of the HS 300 and S&P 500 indices, as of April 2025. The pipeline is designed to maximize temporal coverage while maintaining diversity in model inputs and ensuring comparability across the data sources.

Technical-Chart Images For each stock we segment daily OHLCV data into semi-annual windows and render candlestick charts with matched volume bars. The raw RGB charts are converted to 8-bit grayscale to reduce input dimensionality. We then prompt GPT-4.1 with a fixed instruction to assign one of five long-horizon trend classes—1 (Slightly Up), 2 (Significantly Up), 3 (Flat), 4 (Slightly Down), 5 (Significantly Down)—thereby compressing multi-month dynamics into a single ordinal label that complements subsequent short-term price signals (Figure 2).

Structured Financial Tables For the structured-table modality, we curate a concise yet representative panel of six accounting variables that jointly characterise profitability, liquidity, and capital structure. Specifically, for the HS 300 universe we pull quarterly and annual series from the income statement and cash-flow statement—net profit, operating cash flow, and free cash flow. For the S&P 500 universe we draw analogous series from the balance sheet, cash-flow statement, and statement of changes in equity—shareholders’ equity, operating cash flow, and retained earnings (or accumulated deficit). We align all quarterly and annual financial variables with each firm’s reporting schedule. Period-end financial figures are matched to the closing price on the last trading day of the quarter or year, then forward-filled across all trading days in that period to synchronize with the daily price series.

Price Series and News Text Daily close prices are normalised per stock to enforce stationarity across both markets. After harvesting raw URLs, headlines, and full texts, each article is summarised to 3–4 sentences ($\sim 16\%$ of the original length) using the Sumy latent-semantic-analysis (LSA) algorithm. A relevance weight W_f (Appendix C) biases the summariser toward sentences that mention the focal ticker. To temper the heavy intraday news flow, all same-day summaries for a given stock are aggregated, ranked by ticker-mention frequency, and only the top entry is retained as that day’s representative narrative. The final payload sent to GPT-4.1 never exceeds ten items per request ($temperature = 0$), ensuring deterministic sentiment inference on a 1–5 scale, where 1 denotes negative sentiment and 5 denotes positive sentiment. The resulting scores are finally min–max normalised prior to multimodal fusion.

System: Now you are a financial expert with stock recommendation experience. Based on a specific stock, score for range from 1 to 5, where 1 is negative, 2 is somewhat negative, 3 is neutral, 4 is somewhat positive, 5 is positive. 10 summarized news will be passed in each time, you will give score in format as shown below in the response from assistant.
User: "News to Stock Symbol -- TSLA: Tesla (TSLA) increases production by 22% ### News to Stock Symbol -- TSLA: Tesla (TSLA) faces a 30% drop in deliveries ### News to Stock Symbol -- TSLA: Tesla (TSLA) stock remains stable"
Assistant: "5, 2, 3"
User: "News to Stock Symbol -- TSLA: Tesla (TSLA) unveils new electric vehicle model ### News to Stock Symbol -- TSLA: Tesla (TSLA) faces lawsuit over autopilot feature"
Assistant: "4, 1"
User: ### News to Stock Symbol -- {symbol}: {text}

Figure 2: Prompt–Response Example for Assigning 1–5 Sentiment Scores to News Items

System: Now you are a financial expert analyzing candlestick charts. Based on the candlestick chart provided below, determine the stock price trend. Please output only one of the following numeric values: 1 for Significantly Up, 2 for Slightly Up, 3 for No Change, 4 for Slightly Down, 5 for Significantly Down. 10 gray images will be passed in each time, you will give score in format as shown below in the response from assistant.

User: Images to Stock Symbol -- TSLA: Tesla (TSLA)



Assistant: "1"

User: ### Images to Stock Symbol -- {symbol}: {img}

Figure 3: Prompt–Response Example for Candlestick Chart Six-Month Trend Scoring

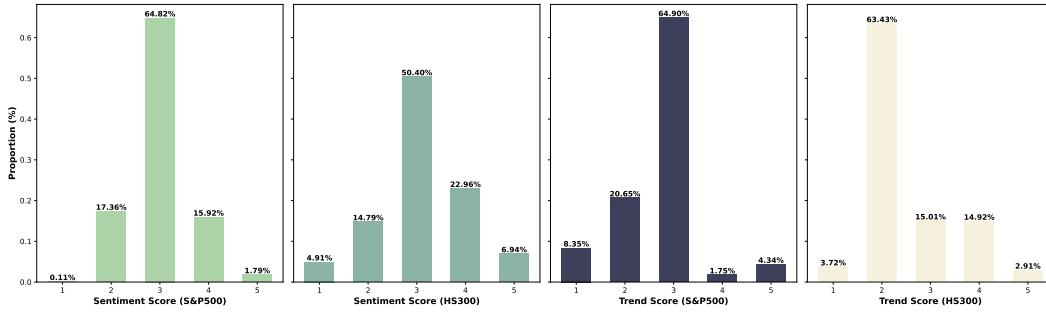


Figure 4: Figures (a) and (b) show the proportions of LSA-generated news sentiment scores (1 = negative, 2 = somewhat negative, 3 = neutral, 4 = somewhat positive, 5 = positive) for S&P 500 and HS300 stocks, respectively. Figures (c) and (d) display the corresponding six-month candlestick-chart trend scores using the same 1–5 scale (1 = negative trend, 5 = positive trend).

Table 5: Chinese (HS300) / English (S&P500) Stock Time Series Data

Date	Open	High	Low	Close	Volume	Dividends	Stock Splits
2025-03-27 00:00:00	11.3700	11.4100	11.3500	11.3900	55334940	0	0
2025-03-28 00:00:00	11.3900	11.4000	11.3400	11.3500	64494555	0.1275	1.2
2025-03-31 00:00:00	11.3600	11.3800	11.2600	11.2600	111612564	0	0
...

In summary, Figures 2 and 3 illustrate the five-level sentiment rubric, while Figure 4 shows that the resulting score distributions are approximately Gaussian. Mild asymmetry is observable: S&P 500 scores lean left (a slight negative bias), whereas HS 300 scores lean right (neutral-to-positive skew), consistent with a gentle U.S. pull-back versus a protracted Chinese rally during the sampling window and with editorial tone differences across English and Chinese outlets.

3 FinMultiTime Property

Upon completion of data mining and preprocessing, FinMultiTime is now primed for analytical evaluation. This section presents the key insights derived from various analytical approaches.

3.1 Dataset Overview

FinMultiTime comprises a comprehensive and heterogeneous dataset totaling over 112.6 GB. Table 5 illustrates representative price time series drawn from the bilingual corpus, while Figure 5 presents corresponding news sentiment scores alongside summaries generated via latent semantic analysis. These multidimensional data points underscore the dataset’s depth and breadth. The assembly process required approximately 5 TB of computing resources over a 60-day period, reflecting our dedication to overcoming challenges and ensuring the robustness of subsequent analyses.

Date	2022-05-02 00:00:00 UTC	2025-03-29 17:22:52
Symbol	\$TSLA (Tesla)	002594.SZ (比亚迪)
Headline	Elon Musk Twitter Co-investors Will Be Rare Birds	王传福：我国新能源汽车技术、产品和产业链均领先全球 3-5 年
Text	When there's a \$44 billion merger in the offing, it's natural that lots of investors kick the tires. With Elon Musk's personal buyout of Twitter, however, it's tough for managers of other people's money – like private equity firms, for example – to justify investing alongside the Tesla chief executive. Morgan Stanley has spearheaded a \$13 billion debt package for the acquisition. It's highly leveraged, at 7 times Twitter's forecast cash flow for next year, using data compiled by Refinitiv. But it's on top of \$21 billion of equity, currently committed by Musk alone, and an additional loan of \$12.5 billion backed by five times that value of Tesla stock, owned by Musk... (483 words)	我国新能源汽车无论是技术产品还是产业链，领先全球大概 3 至 5 年，应把握这个窗口期，坚持开放创新，以更高层次的绿色技术和产品推动更高水平的对外开放，在优势互补和开放合作中出海。”3 月 29 日，比亚迪 (359.200, -0.76, -0.21%) 股份有限公司董事长兼总裁王传福在中国电动汽车百人会论坛（2025）高层论坛上表示... (495 词)
URL	https://www.nasdaq.com/articles/elon-musk-twitter-co-investors-will-be-rare-birds	--
LSA Sum	But Musk isn't necessarily offering co-investment opportunities beyond Twitter, and his commitment to free speech on the platform could play badly with authoritarian governments. If these objections could be overcome, any meaningful equity stake would still be a hefty single outlay even for a large investment institution, never mind for individuals and family offices. - Elon Musk, who has agreed to buy Twitter for \$44 billion, has been inundated with offers from potential equity partners to join him in the deal for the social media group... (114 words)	我国新能源汽车技术、产品和产业链均领先全球 3-5 年... (88 词)

Figure 5: LSA-Generated Summaries of English and Chinese Stock News

Table 6: HS300 vs. S&P 500 — Multimodal Record Counts (35 stocks each)

Semi-annual trend images	Quarterly / annual tables	Daily time-series points	News-sentiment scores
HS300	299,923	1,749	26,467
S&P 500	299,923	2,104	51,235
Total	599,846	3,853	77,702

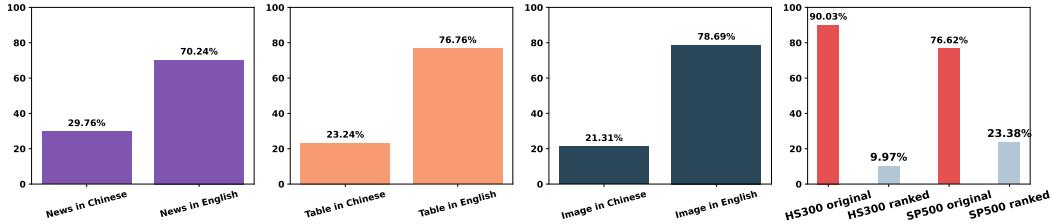


Figure 6: Proportions of Chinese vs. English Modalities (News, Tables, Images) and Coverage Ratios of Ranked vs. Original Daily News for HS300 and S&P 500.

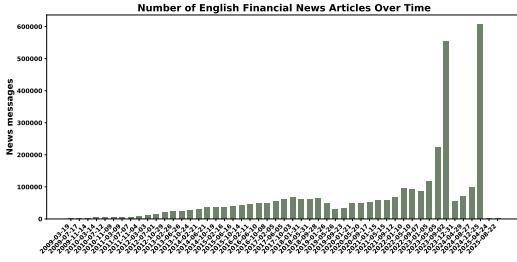


Figure 7

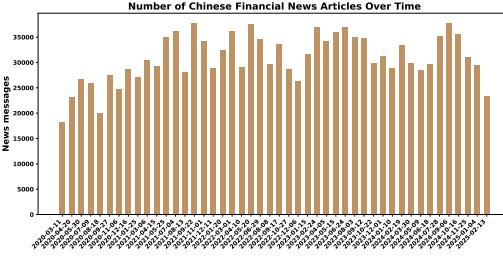


Figure 8

Moreover, we expanded our evaluation to include the 35 most influential constituents of the 2025 S&P 500 and the 35 most influential constituents of the HS300 index—70 stocks in total. These samples were processed through our sentiment-annotation pipeline, yielding 77,702 sentiment-annotated news items, 599,846 semiannual K-line charts, and 3,853 quarterly or annual financial variable records. For detailed information, please refer to Tables 6, 9 and 10 in Appendix D.

3.2 Evaluation

Language Distribution As shown in the first three panels of Figure 6, we compare the proportions of Chinese and English-language news articles, tabular records, and charts to assess FinMultiTime’s multilingual scope and its applicability in a global research context.

Temporal Distribution Figures 8 and 7 plot the annual volume of U.S. stock-market news (1999–2025) and Chinese-market news (2000–2025), respectively. Figures 10 and 9 display the counts of K-line charts for the U.S. market (2006–2025) and the Chinese market (2000–2025),

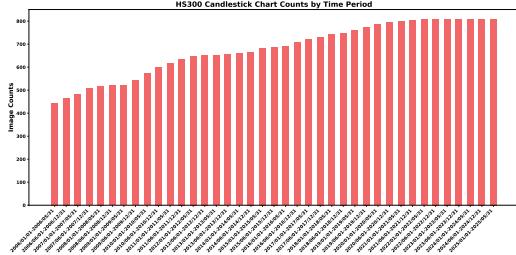


Figure 9

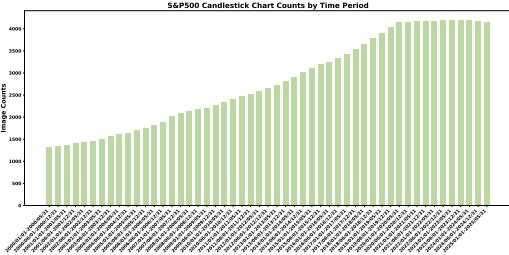


Figure 10

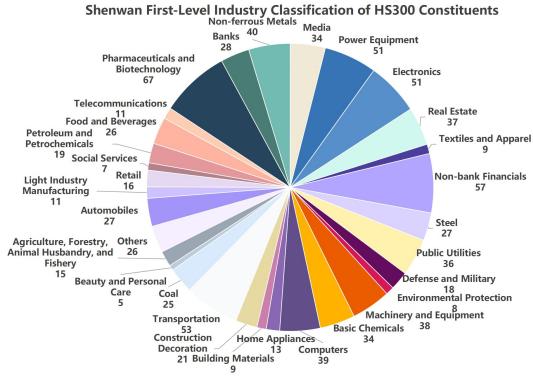


Figure 11

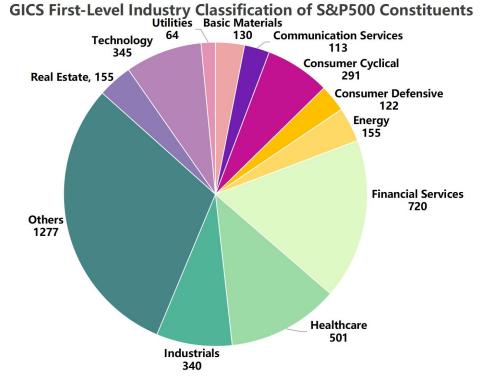


Figure 12

respectively. This temporal analysis reveals evolving trends and patterns, offering valuable insights into the historical progression of financial news coverage.

Industry Distribution Figures 11 and 12 contrast the industry compositions of HS300 constituents (Shenwan Level-1 classification) and S&P 500 constituents (GICS Level-1 classification). Under Shenwan Level-1, HS300 stocks are concentrated in finely segmented sub-sectors—Pharmaceuticals & Biotechnology, Non-Bank Financials, Transportation, Electrical Equipment, and Electronics—whereas the broader GICS Level-1 grouping highlights the dominance of large sectors—Financial Services, Healthcare, Information Technology, and Industrials—among S&P 500 constituents.

Collectively, these analyses demonstrate FinMultiTime's unique value as a benchmark for advanced financial sentiment analysis and time-series forecasting, attributable to its extensive market coverage, robust multilingual support, and deep temporal span.

4 Experiments

To validate the effectiveness of *FinMultiTime*, we conducted both statistical analyses and empirical tests. This section assesses the dataset's overall performance through quantitative and qualitative evaluations. We outline our experimental strategy and demonstrate the dataset's robustness in real-world applications.

4.1 Quantitative Tests

In stock-price forecasting, we use numerical data alongside other modalities—technical K-line charts to capture long-term trends, news articles for market sentiment, and financial statements for company fundamentals to predict short-term price movements. Different models learn different patterns, leading to varied prediction accuracy. We trained on bilingual HS300/S&P 500 datasets of three sizes (5, 15, and 35 stocks) to study how dataset scale affects model performance. We compared six deep-learning architectures:

- **Traditional sequence models:** RNN, LSTM, GRU, and 1D CNN;

Table 7: Model performance across modalities and time horizons for HS 300 stocks. The table compares models on three stock sets (5, 15, and 35 stocks) across four modalities: (1) Time Series Only (price data), (2) News Sentiment (Time Series + Sentiment), (3) Image Trend (Time Series + Long-Term Trend), and (4) Fundamental Table (Time Series + Fundamentals). Performance is measured using MAE, MSE, and R^2 , with lower MAE/MSE and higher R^2 preferred.

#	Model	Time series			News Sentiment			Image Trend			Fundamental Table		
		MAE (↓)	MSE (↓)	R^2 (↑)	MAE (↓)	MSE (↓)	R^2 (↑)	MAE (↓)	MSE (↓)	R^2 (↑)	MAE (↓)	MSE (↓)	R^2 (↑)
5	RNN	0.02234	0.00116	0.82056	0.02776	0.00177	0.79800	0.02513	0.00137	0.79978	0.02716	0.00152	0.76220
	LSTM	0.02542	0.00172	0.83107	0.02080	0.00105	0.86089	0.02079	0.00105	0.85414	0.03132	0.00234	0.73257
	GRU	0.02719	0.00157	0.84827	0.02630	0.00142	0.83780	0.02666	0.00156	0.81649	0.03566	0.00341	0.70141
	CNN	0.04348	0.00366	0.59897	0.04465	0.00410	0.54948	0.04653	0.00458	0.46920	0.03897	0.00315	0.59196
	TimesNet	0.06186	0.00818	0.67671	0.11162	0.01659	0.47544	0.13925	0.02421	0.11394	0.16613	0.03629	0.26917
	Transformer	0.01780	0.00066	0.93733	0.02015	0.00080	0.92080	0.03642	0.00268	0.76037	0.02900	0.00154	0.83630
15	RNN	0.07695	0.01249	0.00002	0.07816	0.01285	0.00005	0.07822	0.01254	0.02675	0.04034	0.00361	0.62207
	LSTM	0.01944	0.00094	0.86228	0.01970	0.00098	0.85608	0.01935	0.00093	0.86178	0.02655	0.00166	0.75171
	GRU	0.02753	0.00156	0.79958	0.02512	0.00143	0.84372	0.02830	0.00172	0.79345	0.02901	0.00245	0.80731
	CNN	0.04140	0.00359	0.57989	0.04405	0.00439	0.56188	0.04442	0.00430	0.50440	0.03906	0.00334	0.63072
	TimesNet	0.15288	0.03262	0.13849	0.15851	0.03482	0.07265	0.12259	0.02328	0.45728	0.20142	0.05210	0.29106
	Transformer	0.01338	0.00036	0.97988	0.01007	0.00027	0.98360	0.01414	0.00048	0.96898	0.02353	0.00111	0.87094
35	RNN	0.07705	0.01236	0.00002	0.03660	0.00358	0.65617	0.05393	0.00542	0.48801	0.05326	0.00605	0.56681
	LSTM	0.01901	0.00093	0.86144	0.01912	0.00092	0.86419	0.01921	0.00093	0.85989	0.02923	0.00289	0.78066
	GRU	0.02454	0.00131	0.85170	0.02557	0.00154	0.83702	0.02909	0.00168	0.80373	0.02640	0.00159	0.83099
	CNN	0.04073	0.00366	0.58944	0.03824	0.00300	0.62581	0.04482	0.00466	0.64729	0.03790	0.00326	0.67346
	TimesNet	0.12852	0.02045	0.31963	0.07271	0.00878	0.60841	0.15302	0.03516	0.00003	0.16498	0.03773	0.30424
	Transformer	0.01511	0.00042	0.97917	0.01224	0.00024	0.98531	0.02420	0.00093	0.95488	0.01550	0.00045	0.96928

Table 8: Model performance across modalities and time horizons for S&P 500 stocks. The table compares models on three stock sets (5, 15, and 35 stocks) across four modalities: (1) Time Series Only (price data), (2) News Sentiment (Time Series + Sentiment), (3) Image Trend (Time Series + Long-Term Trend), and (4) Fundamental Table (Time Series + Fundamentals). Performance is measured using MAE, MSE, and R^2 , with lower MAE/MSE and higher R^2 preferred.

#	Model	Time series			News Sentiment			Image Trend			Fundamental Table		
		MAE (↓)	MSE (↓)	R^2 (↑)	MAE (↓)	MSE (↓)	R^2 (↑)	MAE (↓)	MSE (↓)	R^2 (↑)	MAE (↓)	MSE (↓)	R^2 (↑)
5	RNN	0.05916	0.00794	0.65484	0.03489	0.00333	0.85576	0.02299	0.00110	0.81859	0.05897	0.00768	0.49686
	LSTM	0.02100	0.00087	0.84453	0.01885	0.00073	0.86457	0.02058	0.00085	0.84102	0.01919	0.00071	0.83863
	GRU	0.02444	0.00114	0.82495	0.01999	0.00081	0.88305	0.02575	0.00138	0.84078	0.03015	0.00197	0.73536
	CNN	0.03022	0.00161	0.76389	0.03672	0.00240	0.62313	0.03158	0.00178	0.75333	0.03801	0.00241	0.60509
	TimesNet	0.02896	0.00050	0.74920	0.09056	0.01085	0.95963	0.19695	0.04562	0.67533	0.21783	0.05599	0.71183
	Transformer	0.01553	0.00038	0.90547	0.01570	0.00038	0.91143	0.01554	0.00041	0.91156	0.02548	0.00091	0.75862
15	RNN	0.09155	0.01374	0.00005	0.09376	0.01427	0.00003	0.09121	0.01366	0.00002	0.03844	0.00252	0.65427
	LSTM	0.01743	0.00062	0.87588	0.01821	0.00068	0.87041	0.02000	0.00083	0.85431	0.02478	0.00116	0.89770
	GRU	0.02179	0.00095	0.86501	0.02092	0.00086	0.86255	0.02092	0.00086	0.86827	0.02260	0.00101	0.85699
	CNN	0.02915	0.00152	0.78079	0.03335	0.00189	0.72286	0.03045	0.00164	0.74861	0.03539	0.00209	0.67124
	TimesNet	0.12684	0.02243	0.62065	0.12017	0.01987	0.59290	0.08556	0.01365	0.34835	0.14662	0.02815	0.00002
	Transformer	0.01032	0.00019	0.95211	0.00851	0.00013	0.97187	0.01045	0.00018	0.95556	0.01924	0.00061	0.82294
35	RNN	0.09167	0.01376	0.00003	0.05543	0.00441	0.53614	0.04604	0.00371	0.34601	0.03237	0.00185	0.66103
	LSTM	0.01781	0.00065	0.87352	0.01740	0.00063	0.87564	0.01772	0.00064	0.87262	0.01736	0.00062	0.87730
	GRU	0.01982	0.00079	0.87487	0.02167	0.00091	0.86255	0.02231	0.00100	0.84853	0.02423	0.00126	0.85499
	CNN	0.02780	0.00136	0.79162	0.03231	0.00181	0.75758	0.02983	0.00157	0.77010	0.03318	0.00193	0.71967
	TimesNet	0.12433	0.01753	0.69342	0.07605	0.00779	0.40993	0.22229	0.05952	0.71709	0.12692	0.02098	0.00002
	Transformer	0.00477	0.00005	0.98959	0.00659	0.00008	0.98110	0.00888	0.00012	0.97247	0.00887	0.00012	0.97357

- **Recent time-series methods:** 4-layer Vanilla Transformer and 4-layer TimesNet.

All experiments used 50 days of historical data to forecast the next 3 days, training each model for 100 epochs. We evaluated on the 5-stock split, removed one outlier, and reported the mean results.

4.1.1 Test Results on HS300

Results on HS300 are shown in table 7. As the training set grew from 5 to 35 stocks, we found that *Transformer* achieved the highest accuracy (average $R^2 \approx 0.97$ at 35 stocks); *LSTM* ranked second ($R^2 \approx 0.84$); *GRU* ranked third ($R^2 \approx 0.83$); *TimesNet* performed worst ($R^2 \approx 0.31$).

The low performance of basic sequence models can be due to the problem of signal amplification for weak learners. RNNs and basic sequence models struggle to fully capture fundamental ratios or alternative data if trained alone. Merging modalities injects complementary views which are hard to capture for basic models. The reason why Transformer and LSTM sometimes not working

well can due to the diminishing returns for strong learners. Models like Transformers or LSTMs already achieve about 0.95 R^2 on single streams. Adding noisy, lower-quality modalities can actually introduce variance and drag down accuracy.

These results confirm FinMultiTime’s effectiveness and robustness for financial modeling and sentiment analysis: larger, multimodal training sets yield substantial gains, while small datasets are inherently limited.

4.1.2 Test Results on S&P 500

As shown in the Table 8, the U.S. S&P 500 split shows a nearly identical pattern as HS300. The results shows that *Transformer* again leads with $R^2 \approx 0.97$ at 35 stocks; *LSTM* and *GRU* follow ($R^2 \approx 0.84$ and 0.83); *TimesNet* remains last ($R^2 \approx 0.31$).

These results again confirm that larger, multimodal training sets yield substantial gains. This cross-market consistency underlines the dataset’s general utility.

4.2 Sentiment Effectiveness

In the Table 7 and Table 8, only Transformer and LSTM models consistently benefited from adding sentiment, trend, or fundamental inputs; GRU saw only occasional gains. RNN, CNN, and TimesNet often treated extra modalities as noise. Interestingly, on the smallest training set (5 stocks), LSTM slightly outperformed Transformer, but as data volume grew, Transformer’s accuracy advantage became pronounced.

4.3 Discussion

Hyperparameter tuning can affect these results, but to ensure fair comparison we held most settings constant, which may have constrained some models. We encoded sentiment and trend with 1–5 scores, a granularity that may omit nuance. Prior studies report strong impacts of news, trend, or fundamentals on prices, yet our gains were modest due to two factors:

1. Our models already achieved high baseline accuracy, leaving little headroom;
2. Delays in news propagation and the inability of past trends or static fundamentals to capture unforeseen shocks can limit immediate predictive value.

5 Related Work

Financial time-series models Traditional time-series models like linear regression [30], ARIMA [2] and GARCH [3] depend on stationarity and strong assumptions, so they often miss complex dependencies or abrupt shocks. Recently, machine learning [17, 12, 13], deep learning [27] and NLP [34, 24] have tapped sentiment and other qualitative signals to enhance forecast accuracy. This trend mirrors Markowitz’s market correlation concept, linking sentiment from news, blogs, and social media to asset prices. With growing data and compute, LLMs now enable finer sentiment quantification [23]. Moreover, TSMixer-MICM [16] turns quarterly financial-statement tables into time-series features, aligning them with price and text data for three-modal analysis.

6 Conclusion

Based on the results of the FNSPID experiments, we derive three primary conclusions that contribute to the understanding of stock-price forecasting using deep learning techniques. First, the quality and scale of the dataset play a pivotal role in determining the accuracy of predictive models. Larger and more refined datasets provide richer context and reduce noise, thereby enabling models to learn more robust representations of market dynamics. Second, the integration of high-quality multimodal inputs—such as combining textual news data, numerical stock indicators, and technical signals—substantially enhances the performance of Transformer-based architectures. This multimodal fusion allows the model to capture complex inter-dependencies that single-modality inputs often overlook. Finally, Transformer models demonstrate clear superiority over both traditional

time-series forecasting methods (e.g., ARIMA, LSTM) and recent state-of-the-art approaches such as TimesNet. This highlights the efficacy of attention mechanisms in modeling temporal dependencies and long-range correlations within financial time-series data.

References

- [1] Dogu Araci. Finbert: Financial sentiment analysis with pre-trained language models. *arXiv preprint arXiv:1908.10063*, 2019.
- [2] Adebiyi A Ariyo, Adewumi O Adewumi, and Charles K Ayo. Stock price prediction using the arima model. In *2014 UKSim-AMSS 16th international conference on computer modelling and simulation*, pages 106–112. IEEE, 2014.
- [3] Luc Bauwens, Sébastien Laurent, and Jeroen VK Rombouts. Multivariate garch models: a survey. *Journal of applied econometrics*, 21(1):79–109, 2006.
- [4] Lei Chai, Hongfeng Xu, Zhiming Luo, and Shaozi Li. A multi-source heterogeneous data analytic method for future price fluctuation prediction. *Neurocomputing*, 418:11–20, 2020.
- [5] Clayton Leroy Chapman, Lars Hillebrand, Marc Robin Stenzel, Tobias Deußer, David Biesner, Christian Bauckhage, and Rafet Sifa. Towards generating financial reports from tabular data using transformers. In *International Cross-Domain Conference for Machine Learning and Knowledge Extraction*, pages 221–232. Springer, 2022.
- [6] Zihan Chen, Lei Nico Zheng, Cheng Lu, Jialu Yuan, and Di Zhu. Chatgpt informed graph neural network for stock movement prediction. *arXiv preprint arXiv:2306.03763*, 2023.
- [7] Junyan Cheng and Peter Chin. Sociodojo: Building lifelong analytical agents with real-world text and time series. In *The Twelfth International Conference on Learning Representations*, 2024.
- [8] Zihan Dong, Xinyu Fan, and Zhiyuan Peng. Fnspid: A comprehensive financial news dataset in time series. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 4918–4927, 2024.
- [9] Kui Fu and Yanbin Zhang. Incorporating multi-source market sentiment and price data for stock price prediction. *Mathematics*, 12(10):1572, 2024.
- [10] Udit Gupta. Gpt-investar: Enhancing stock investment strategies through annual report analysis with large language models. *arXiv preprint arXiv:2309.03079*, 2023.
- [11] Yu Huang, Chenzhuang Du, Zihui Xue, Xuanyao Chen, Hang Zhao, and Longbo Huang. What makes multi-modal learning better than single (provably). *Advances in Neural Information Processing Systems*, 34:10944–10956, 2021.
- [12] Bryan Kelly, Dacheng Xiu, et al. Financial machine learning. *Foundations and Trends® in Finance*, 13(3-4):205–363, 2023.
- [13] Kyoung-jae Kim. Financial time series forecasting using support vector machines. *Neurocomputing*, 55(1-2):307–319, 2003.
- [14] Kelvin JL Koa, Yunshan Ma, Ritchie Ng, and Tat-Seng Chua. Learning to generate explainable stock predictions using self-reflective large language models. In *Proceedings of the ACM Web Conference 2024*, pages 4304–4315, 2024.
- [15] Yaxuan Kong, Yiyuan Yang, Yoontae Hwang, Wenjie Du, Stefan Zohren, Zhangyang Wang, Ming Jin, and Qingsong Wen. Time-mqa: Time series multi-task question answering with context enhancement. *arXiv preprint arXiv:2503.01875*, 2025.
- [16] Ross Koval, Nicholas Andrews, and Xifeng Yan. Financial forecasting from textual and tabular time series. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 8289–8300, 2024.
- [17] Bjoern Krollner, Bruce Vanstone, and Gavin Finnie. Financial time series forecasting with machine learning techniques: A survey. In *European Symposium on Artificial Neural Networks: Computational Intelligence and Machine Learning*, pages 25–30, 2010.

- [18] Moreno La Quatra and Luca Cagliero. End-to-end training for financial report summarization. In *Proceedings of the 1st Joint Workshop on Financial Narrative Processing and MultiLing Financial Summarisation*, pages 118–123, 2020.
- [19] Geon Lee, Wenchao Yu, Kijung Shin, Wei Cheng, and Haifeng Chen. Timecap: Learning to contextualize, augment, and predict time series events with large language model agents. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pages 18082–18090, 2025.
- [20] Haoxin Liu, Shangqing Xu, Zhiyuan Zhao, Lingkai Kong, Harshavardhan Prabhakar Kamarthi, Aditya Sasanur, Megha Sharma, Jiaming Cui, Qingsong Wen, Chao Zhang, et al. Time-mmd: Multi-domain multimodal dataset for time series analysis. *Advances in Neural Information Processing Systems*, 37:77888–77933, 2024.
- [21] Xiao-Yang Liu, Guoxuan Wang, Hongyang Yang, and Daochen Zha. Fingpt: Democratizing internet-scale data for financial large language models. In *NeurIPS 2023 Workshop on Instruction Tuning and Instruction Following*.
- [22] Xiao-Yang Liu, Hongyang Yang, Qian Chen, Runjia Zhang, Liuqing Yang, Bowen Xiao, and Christina Dan Wang. Finrl: A deep reinforcement learning library for automated stock trading in quantitative finance. *arXiv preprint arXiv:2011.09607*, 2020.
- [23] Alejandro Lopez-Lira and Yuehua Tang. Can chatgpt forecast stock price movements? return predictability and large language models. *arXiv preprint arXiv:2304.07619*, 2023.
- [24] Mantas Lukauskas, Vaida Pilinkienė, Jurgita Bruneckienė, Alina Stundžienė, Andrius Grybauskas, and Tomas Ruzgas. Economic activity forecasting based on the sentiment analysis of news. *Mathematics*, 10(19):3461, 2022.
- [25] Burton G. Malkiel and Eugene F. Fama. Efficient capital markets: A review of theory and empirical work. *The Journal of Finance*, 25(2):383–417, 1970.
- [26] Eliza Mik. Smart contracts: terminology, technical limitations and real world complexity. *Law, innovation and technology*, 9(2):269–300, 2017.
- [27] Omer Berat Sezer, Mehmet Ugur Gudelek, and Ahmet Murat Ozbayoglu. Financial time series forecasting with deep learning: A systematic literature review: 2005–2019. *Applied soft computing*, 90:106181, 2020.
- [28] Wataru Souma, Irena Vodenska, and Hideaki Aoyama. Enhanced news sentiment analysis using deep learning methods. *Journal of Computational Social Science*, 2(1):33–46, 2019.
- [29] Xinlei Wang, Maike Feng, Jing Qiu, Jinjin Gu, and Junhua Zhao. From news to forecast: Integrating event analysis in llm-based time series forecasting with reflection. *Advances in Neural Information Processing Systems*, 37:58118–58153, 2024.
- [30] Sanford Weisberg. *Applied linear regression*, volume 528. John Wiley & Sons, 2005.
- [31] Andrew Robert Williams, Arjun Ashok, Étienne Marcotte, Valentina Zantedeschi, Jithendaraa Subramanian, Roland Riachi, James Requeima, Alexandre Lacoste, Irina Rish, Nicolas Chapados, et al. Context is key: A benchmark for forecasting with essential textual information. *arXiv preprint arXiv:2410.18959*, 2024.
- [32] Huizhe Wu, Wei Zhang, Weiwei Shen, and Jun Wang. Hybrid deep sequential modeling for social text-driven stock prediction. In *Proceedings of the 27th ACM international conference on information and knowledge management*, pages 1627–1630, 2018.
- [33] Qianqian Xie, Weiguang Han, Zhengyu Chen, Ruoyu Xiang, Xiao Zhang, Yueru He, Mengxi Xiao, Dong Li, Yongfu Dai, Duanyu Feng, et al. Finben: A holistic financial benchmark for large language models. *Advances in Neural Information Processing Systems*, 37:95716–95743, 2024.
- [34] Frank Z Xing, Erik Cambria, and Roy E Welsch. Natural language based financial forecasting: a survey. *Artificial Intelligence Review*, 50(1):49–73, 2018.

- [35] Yumo Xu and Shay B Cohen. Stock movement prediction from tweets and historical prices. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1970–1979, 2018.
- [36] Hang Yang, Yubo Chen, Kang Liu, Yang Xiao, and Jun Zhao. Dcfee: A document-level chinese financial event extraction system based on automatically labeled training data. In *Proceedings of ACL 2018, System Demonstrations*, pages 50–55, 2018.
- [37] Yi Yang, Yixuan Tang, and Kar Yan Tam. Investlm: A large language model for investment using financial domain instruction tuning. *arXiv preprint arXiv:2309.13064*, 2023.
- [38] Wentao Zhang, Lingxuan Zhao, Haochong Xia, Shuo Sun, Jiaze Sun, Molei Qin, Xinyi Li, Yuqing Zhao, Yilei Zhao, Xinyu Cai, et al. A multimodal foundation agent for financial trading: Tool-augmented, diversified, and generalist. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 4314–4325, 2024.
- [39] Yanzhao Zou and Dorien Herremans. Prebit—a multimodal model with twitter finbert embeddings for extreme price movement prediction of bitcoin. *Expert Systems with Applications*, 233:120838, 2023.

A Related Work

Financial Multimodal time-series datasets Financial Multimodal time-series datasets fall into two groups. General economic collections (e.g., Time-MMD [20], CiK [31]) pair macro-text with monthly indicators but are too coarse and small for fine-grained forecasting. Financial-specific sets target asset prices: NewsForecast links Bitcoin news to daily prices; TimeCAP[19], DOW30[6], TSQA[15] align stock news with prices; ACL18[35], CIKM18[32], SEP[14] use tweet sentiment. FinBEN [33] and FNSPID Nasdaq [8] add bilingual text yet remain text-price only, while a 2024 EMNLP Findings study[16] is first to fuse quarterly tables with text and prices, albeit at low frequency and small scale. Overall, these datasets are modest in size and mostly single-market (chiefly U.S.), limiting their usefulness for pre-training and evaluating emerging large-scale financial LLMs and multimodal models.

B Illustration of the Temporal Distribution

C Bilingual News Summarize Algorithm

In reference to FNSPID [8], we introduce a weight model W_z to enhance summarization and emphasize relevant stocks. In the sumy package, all terms are included in the summary. Exclusiveness involves rephrasing sentences rather than extracting terms. We parse the graph G into sentences and assign a weight W_p based on relevance to the stock symbol, setting $k = 1$ for sentences containing the symbol. For summarized sentences S_{sum} , a score of $t = 1$ is given if the sentence is longer. In Equation (6), we combine W_p and W_q to calculate the final weight W_z , with irrelevant sentences receiving a weight of 0. The sentences are sorted by weight to form the final summary.

$$W_p(S, s) = \begin{cases} k & \text{if } S \in G \\ 0 & \text{otherwise} \end{cases}$$

$$W_q(S_{sum}, S_{long}) = \begin{cases} t & \text{if } S_{sum} \in S_{long} \\ 0 & \text{otherwise} \end{cases}$$

$$W_z = W_p + W_q$$

D Introduction to the Experimental Stock Set

The following tables provide information on 35 s&P500 /HS 300 stocks, listing their corresponding sectors and the availability of data in different formats, including image, text, table, and time series. Each stock’s data availability is marked with a check symbol for each format.

Table 9: 35 HS300 stock Information

Stock Symbol	Sector	Image	Text	Table	Time Series
002371.SZ	Semiconductor	✓	✓	✓	✓
601318.SH	Insurance	✓	✓	✓	✓
300750.SZ	Battery	✓	✓	✓	✓
600900.SH	Power Industry	✓	✓	✓	✓
300124.SZ	Electronic Components	✓	✓	✓	✓
600031.SH	Construction Machinery	✓	✓	✓	✓
300274.SZ	Photovoltaic Equipment	✓	✓	✓	✓
000725.SZ	Optoelectronics	✓	✓	✓	✓
300059.SZ	Internet Services	✓	✓	✓	✓
600309.SH	Chemical Products	✓	✓	✓	✓
600276.SH	Pharmaceutical	✓	✓	✓	✓
002415.SZ	Computer Equipment	✓	✓	✓	✓
000333.SZ		✓	✓	✓	✓
000651.SZ	Home Appliances	✓	✓	✓	✓
600690.SH		✓	✓	✓	✓
601088.SH	Coal Industry	✓	✓	✓	✓
600519.SH	Liquor	✓	✓	✓	✓
600809.SH		✓	✓	✓	✓
002714.SZ	Agriculture	✓	✓	✓	✓
002594.SZ	Automobile	✓	✓	✓	✓
601127.SH		✓	✓	✓	✓
600887.SH	Food	✓	✓	✓	✓
000063.SZ	Communication Equipment	✓	✓	✓	✓
002352.SZ	Logistics	✓	✓	✓	✓
002475.SZ	Consumer Electronics	✓	✓	✓	✓
300760.SZ	Medical Devices	✓	✓	✓	✓
600036.SH		✓	✓	✓	✓
601166.SH		✓	✓	✓	✓
601288.SH		✓	✓	✓	✓
600919.SH	Banking	✓	✓	✓	✓
600000.SH		✓	✓	✓	✓
000001.SZ		✓	✓	✓	✓
601229.SH		✓	✓	✓	✓
600030.SH	Securities	✓	✓	✓	✓
601211.SH		✓	✓	✓	✓

E FinMultiTime Applications

This section critically examines the potential uses of the FinMultiTime dataset in financial-market research, the technical hurdles encountered during its construction, and the attendant ethical challenges, while outlining avenues for future work.

E.1 Construction Challenges

Bilingual news extraction and sentiment labelling. We experimented with lightweight extractive algorithms (Luhn, LexRank, TextRank) and generative models (distilbart-cnn-12-6). Although both approaches handle simple sentiments (e.g., “sharp price rise” or “steep decline”) reasonably well, extractive methods often miss key context in longer passages, whereas generative models suffer from summary repetition, unstable scores, and attention drift on lengthy documents.

Table 10: 35 S&P500 Stock Information

Stock Symbol	Sector	Image	Time Series	Text	Table
GOOG	Communication Services	✓	✓	✓	✓
DIS		✓	✓	✓	✓
BKNG	Consumer Cyclical	✓	✓	✓	✓
TJX		✓	✓	✓	✓
COST		✓	✓	✓	✓
KO	Consumer Defensive	✓	✓	✓	✓
PM		✓	✓	✓	✓
PEP		✓	✓	✓	✓
XOM	Energy	✓	✓	✓	✓
CVX		✓	✓	✓	✓
V		✓	✓	✓	✗
WFC		✓	✓	✓	✓
GS	Financial Services	✓	✓	✓	✓
PGR		✓	✓	✓	✓
MS		✓	✓	✓	✓
ABBV		✓	✓	✓	✓
ABT		✓	✓	✓	✓
MRK	Healthcare	✓	✓	✓	✓
TMO		✓	✓	✓	✓
BSX		✓	✓	✓	✓
AMGN		✓	✓	✓	✓
GE		✓	✓	✓	✓
BA	Industrials	✓	✓	✓	✓
UNP		✓	✓	✓	✓
AAPL		✓	✓	✓	✓
NVDA		✓	✓	✓	✓
CRM		✓	✓	✓	✓
ORCL		✓	✓	✓	✓
NOW	Technology	✓	✓	✓	✓
ACN		✓	✓	✓	✓
ADBE		✓	✓	✓	✓
AMD		✓	✓	✓	✓
QCOM		✓	✓	✓	✓
TXN		✓	✓	✓	✓
NEE	Utilities	✓	✓	✓	✓

Modal imbalance. Relying on a small set of tabular variables or on trend labels derived solely from candlestick images fails to unlock the complementary value of FinMultiTime’s four modalities. These limitations underscore the need for more efficient architectures that can exploit mutual information among text, images, and structured data to reveal genuine predictive power.

E.2 Prospective Use Cases

Multimodal model training. The temporally aligned text, numeric, image, and table streams enable the development of joint-learning models for stock prediction. Such models can bolster robustness to short-horizon noise and improve reinforcement-learning agents in sequential decision-making—especially for trend forecasting and strategy design.

Sentiment and trend-signal analysis. Combining news-sentiment scores with long-horizon trend labels allows researchers to assess the incremental explanatory power of non-price signals within a

modern portfolio-theory framework. Batch processing of sentiment and trend indicators across many tickers further refines market forecasts and portfolio allocation.

Correlation and anomaly detection. The four aligned modalities facilitate granular studies of how sentiment, image-based trends, and fundamentals correlate with price dynamics, potentially revealing latent market drivers. Pattern matching on historical data can surface precursors of systemic risk, offering fresh tools for volatility warnings and risk management.

Financial generative-AI applications. With its large, heterogeneous corpus, FinMultiTime serves as prime fine-tuning material for large language models, powering next-generation robo-advisers, automated report writers, and other finance-oriented AI services.

E.3 Ethical Considerations

Privacy and data security. Financial records often contain sensitive personal or institutional information. We employ state-of-the-art anonymisation and de-identification techniques and adhere strictly to GDPR, CCPA, and related regulations to safeguard privacy throughout data collection and processing.

Misuse risks. Predictive models built on FinMultiTime could be misappropriated, leading to market manipulation or systemic risk. We therefore conduct bias and fairness audits and publish explicit usage guidelines to curb discriminatory or misleading outcomes.

Transparency and traceability. Every record is source-tagged, and detailed processing documentation is released publicly, ensuring reproducibility, auditability, and responsible research practice.

By addressing construction bottlenecks, enriching multimodal use cases, and enforcing rigorous ethical safeguards, FinMultiTime not only provides a solid empirical foundation for financial-market analysis but also sets a high academic and ethical benchmark for future industry and scholarly endeavours.

F Future Work

Expanding the FinMultiTime Dataset: Although our coverage of stock-related data is extensive, financial data remain inherently time-sensitive. We plan to develop an automated pipeline to continuously ingest and update news feeds, thereby substantially enlarging the dataset’s scope and currency.

Unlocking FinMultiTime’s Full Potential: As the most comprehensive resource aligning price series, sentiment annotations, long-term trend signals, and corporate fundamental data, FinMultiTime can support several frontier research directions:

Multimodal Modeling: Multimodal modeling will integrate heterogeneous sources—text, images, tables, and time series—to construct more robust market-prediction models; sentiment-impact analysis will quantitatively assess how news sentiment drives stock-price volatility, thereby advancing sentiment-analysis algorithms; trend-signal evaluation will investigate the contribution of long-term trend indicators to forecasting accuracy; and fundamental-data integration will examine the auxiliary role of financial-statement features in investment decision-making to enhance real-world applicability. Although our news coverage is already extensive, the synergistic exploitation of chart images, textual summaries, and tabular data remains underexploited. In future work, we will explore pre-training language models within a reinforcement-learning framework to improve multimodal feature extraction and its downstream applications.

By identifying these limitations and outlining targeted research avenues, we aim to inspire subsequent studies and further enhance the value and impact of the FinMultiTime dataset.

NeurIPS Paper Checklist

1. Claims

Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

Answer: [Yes]

Justification: The main claims made in the abstract and introduction accurately reflect the paper's contributions and scope.

Guidelines:

- The answer NA means that the abstract and introduction do not include the claims made in the paper.
- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. Limitations

Question: Does the paper discuss the limitations of the work performed by the authors?

Answer: [Yes]

Justification: The limitations of the work are discussed and the future directions are also provided.

Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. Theory assumptions and proofs

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [NA]

Justification: The paper contains no theoretical results.

Guidelines:

- The answer NA means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. Experimental result reproducibility

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: The details about experiments are provided.

Guidelines:

- The answer NA means that the paper does not include experiments.
- If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general, releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
 - (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
 - (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
 - (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
 - (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. Open access to data and code

Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [Yes]

Justification: The code and its documents of this research have been released at <https://github.com/Marigoldwu/PyDGC>.

Guidelines:

- The answer NA means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so “No” is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (<https://nips.cc/public/guides/CodeSubmissionPolicy>) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. Experimental setting/details

Question: Does the paper specify all the training and test details (e.g., data splits, hyper-parameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [Yes]

Justification: The paper specifies all the training details.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. Experiment statistical significance

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [Yes]

Justification: The main experimental results contain appropriate information about the statistical significance.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. Experiments compute resources

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [\[Yes\]](#)

Justification: The experimental environment is presented.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. Code of ethics

Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics <https://neurips.cc/public/EthicsGuidelines>?

Answer: [\[Yes\]](#)

Justification: We have reviewed the NeurIPS Code of Ethics and confirm that our paper complies fully with all its guidelines.

Guidelines:

- The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
- If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. Broader impacts

Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [\[Yes\]](#)

Justification: We discussed the future work.

Guidelines:

- The answer NA means that there is no societal impact of the work performed.
- If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out

that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.

- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. Safeguards

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

Answer: [NA]

Justification: The paper does not involve any datasets or models that could potentially be misused or present significant ethical risks.

Guidelines:

- The answer NA means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. Licenses for existing assets

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: All third-party assets used in the paper, including code, datasets, and pretrained models, are properly credited by citing the original sources.

Guidelines:

- The answer NA means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, paperswithcode.com/datasets has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. New assets

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: The new assets introduced in the paper are well documented, and the documentation is provided.

Guidelines:

- The answer NA means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. Crowdsourcing and research with human subjects

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [NA]

Justification: This paper does not involve any crowdsourcing experiments or research with human subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. Institutional review board (IRB) approvals or equivalent for research with human subjects

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [NA]

Justification: This study does not involve human participants or subjects.

Guidelines:

- The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. Declaration of LLM usage

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorosity, or originality of the research, declaration is not required.

Answer: [NA]

Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components.

Guidelines:

- The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy (<https://neurips.cc/Conferences/2025/LLM>) for what should or should not be described.