A Comprehensive Review of Multimodal Financial Stock Prediction Research

Li Su

Chengdu University of Information Technology

Received: May 25, 2026

Revised: May 26, 2026

Accepted: May 27, 2026

Published online: May 28, 2026

To appear in: *International Journal of Advanced AI Applications,* Vol.2, No. 6 (June 2026)

* Corresponding Author: Li Su (suli3607@foxmail.com)

Abstract. With the development of financial text mining, deep time-series modeling, and large language models, multimodal stock prediction has evolved from early shallow fusion between price sequences and sentiment signals into an integrated research paradigm covering news text, technical indicators, ESG information, exogenous events, company relation graphs, and long-document semantic compression. Centered on five recent representative studies and supplemented by related work on financial text representation learning, graph neural networks, Transformer-based time-series modeling, and financial large language models, this review synthesizes the main research trajectory, core methodological categories, and key challenges in multimodal financial stock prediction. Specifically, the paper first examines the development of financial text modeling and early bimodal forecasting frameworks, then summarizes the extension of ESG signals, industry context, and exogenous events to stock prediction, and next reviews representative approaches to graph-based relation modeling and stable fusion. It further discusses recent advances in Transformers, patch-based modeling, and large language models for long-text processing and cross-modal alignment. On this basis, the review identifies common issues in label granularity, modality conflict, long-text redundancy, explicit event modeling, and interpretability, and concludes by summarizing the major research trends already visible in the literature.

Keywords: *Multimodal Finance; Stock Prediction; Financial Text; ESG; Event-driven Modeling; Graph Relations; Large Language Models.*

1. Introduction

The stock market is a typical complex system whose price movements are shaped not only by historical trading behavior but also by news sentiment, investor mood, industry conditions, macro policies, and corporate governance. Under this setting, prediction based solely on a single price series often fails to provide sufficient explanatory power for nonlinear market fluctuations and sudden changes.

Early studies focused primarily on the relationship between market sentiment and price movements. One study showed that social media mood has a statistically measurable relationship with market fluctuations [2], while another demonstrated the effectiveness of LSTM models for financial market prediction [8]. As financial text processing has advanced, researchers have gradually integrated various sources of information—such as news, announcements, social media, ESG themes, and event details—with price series. This evolution has facilitated the transition of stock prediction from a unimodal to a multimodal paradigm, reflecting the increasing complexity of factors influencing market behavior.

Within this evolution, several representative contributions have emerged from various perspectives, including deep fusion, ESG enhancement, hierarchical LLM-based text processing, industry-specific scenario extension, and stable multimodal fusion [3, 13, 25, 28, 30]. Together, these studies trace the main developmental trajectory of multimodal financial stock prediction and reflect a significant shift from merely questioning whether multimodal fusion should be utilized to actively examining how it can be effectively implemented. This progression indicates a growing sophistication in the methodologies employed within the field.

Accordingly, this review addresses three questions: how the major modality types and their functional roles have evolved; what progress has been made in text modeling, time-series modeling, and cross-modal fusion across different methodological routes; and which key challenges and research trends remain central in the field.

2. Foundations of Multimodal Stock Prediction

2.1. Financial Text Representation and Sentiment Modeling

Financial text is one of the earliest and most widely used sources of unstructured information in multimodal stock prediction. One study demonstrated through semantic orientation analysis of economic texts that sentiment information within the financial domain can be modeled in a stable manner [18]. At the dataset level, the FiQA challenge at WWW 2018 played a crucial role in promoting the development of financial opinion mining and question answering, thereby providing an important data foundation for subsequent research on financial sentiment analysis and aspect-level language understanding [17]. This progression underscores the significance of leveraging textual data to enhance predictive modeling in finance.

Among pretrained language models, FinBERT [1] marks a major milestone. By further training BERT on financial-domain corpora, FinBERT substantially improves semantic representation for financial news, announcements, and market commentary. With the emergence of financial large language models such as FinGPT [27], financial text representation has expanded from sentiment classification to more complex tasks, including summarization, question answering, explanation generation, and long-document reasoning.

2.2. Financial Time-Series Modeling

Deep time-series modeling constitutes another core line of research in stock prediction. It provided early evidence that LSTM has competitive predictive power in financial markets [8], after which LSTM, BiLSTM, GRU, and their attention-based extensions were widely used for stock movement prediction. With the transition into the Transformer era, the original Transformer [23] offered a unified architecture for long-range dependency modeling, while PatchTST [19], TimesNet [29], and iTransformer [15] further improved long-sequence modeling efficiency and generalization.

These developments in time-series modeling have provided a strong numerical backbone for multimodal financial research. As temporal modeling capacity has improved, researchers have been able to combine prices more effectively with heterogeneous information such as text, events, and graph relations.

3. From Bimodal to Multimodal: Evolution of the Research Trajectory

3.1. Formation of Text-Price Bimodal Frameworks

The earliest mature form of multimodal stock prediction was predominantly based on joint modeling of text and price. It proposed StockNet, which uses tweet text together with historical prices for stock movement prediction and remains a key early study in publicly accessible evaluation settings [26]. Building on this line, further combined social-media text with company relations [21], showing that textual information and market structure are not independent of each other.

This line of research was advanced by proposing a deep fusion model that combines a FinBERT text branch with an LSTM time-series branch to jointly model prices, technical indicators, and news headlines. The work clarified the now-familiar architecture consisting of a domain-specific text encoder, a temporal encoder, and a fusion layer, which has made bimodal frameworks both more reproducible and extensible [3]. This innovative approach emphasizes the potential of integrating distinct data types for improved predictive performance in financial applications.

As shown in Figure 1, a FinBERT-based text branch and an LSTM-based time-series branch are mapped into a unified fusion layer to generate stock trend predictions. The present image is consistent with a dual-branch deep fusion architecture that focuses on integrating text, price series, and technical indicators [3]. This approach highlights the effectiveness of combining various data modalities for improved accuracy in forecasting stock trends.

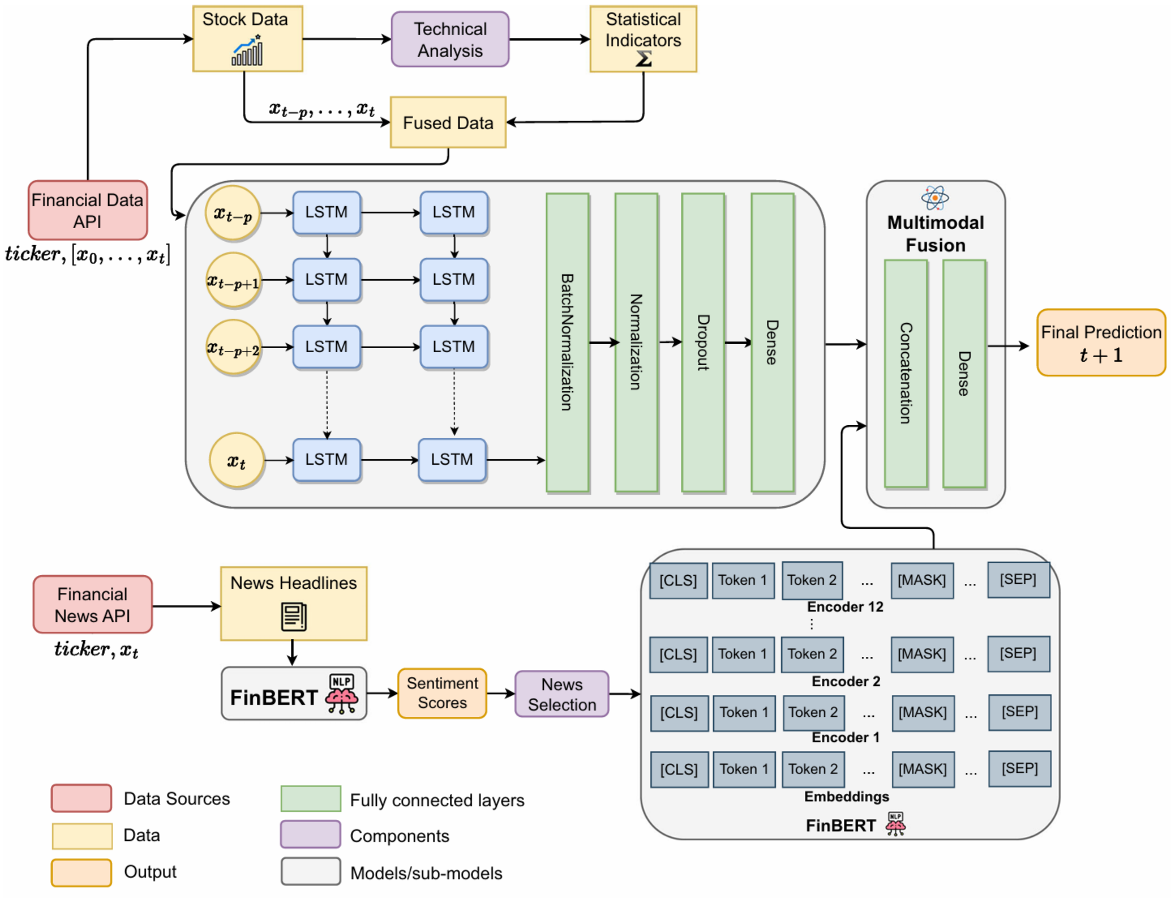

Figure 1. Deep fusion architecture for multimodal stock prediction (adapted from [3]).

At this stage, the primary objective was to validate the complementarity between financial text and historical prices, with the central contribution being the elevation of text from merely an auxiliary explanatory variable to a recognized formal input modality. However, despite this advancement, most methods in this phase continued to rely primarily on short text or news headlines, failing to adequately address challenges such as long-text redundancy, news sparsity, and cross-modal conflicts. This limitation highlights the need for more robust techniques that can effectively utilize longer texts and diverse data sources to enhance overall predictive accuracy. As researchers continue to explore these issues, the potential for improved stock prediction models remains significant.

3.2. ESG Information and Industry-Scenario Extension

As bimodal frameworks gradually matured, researchers began introducing signals with stronger medium- and long-term explanatory value. One significant study combined ESG news sentiment with technical indicators to predict the S&P 500 index, demonstrating that ESG variables not only reflect aspects of corporate governance and sustainability but are also closely associated with investor expectations and market pricing [13]. This integration emphasizes the growing recognition of ESG factors as vital components in financial analysis and highlights their influence on market dynamics.

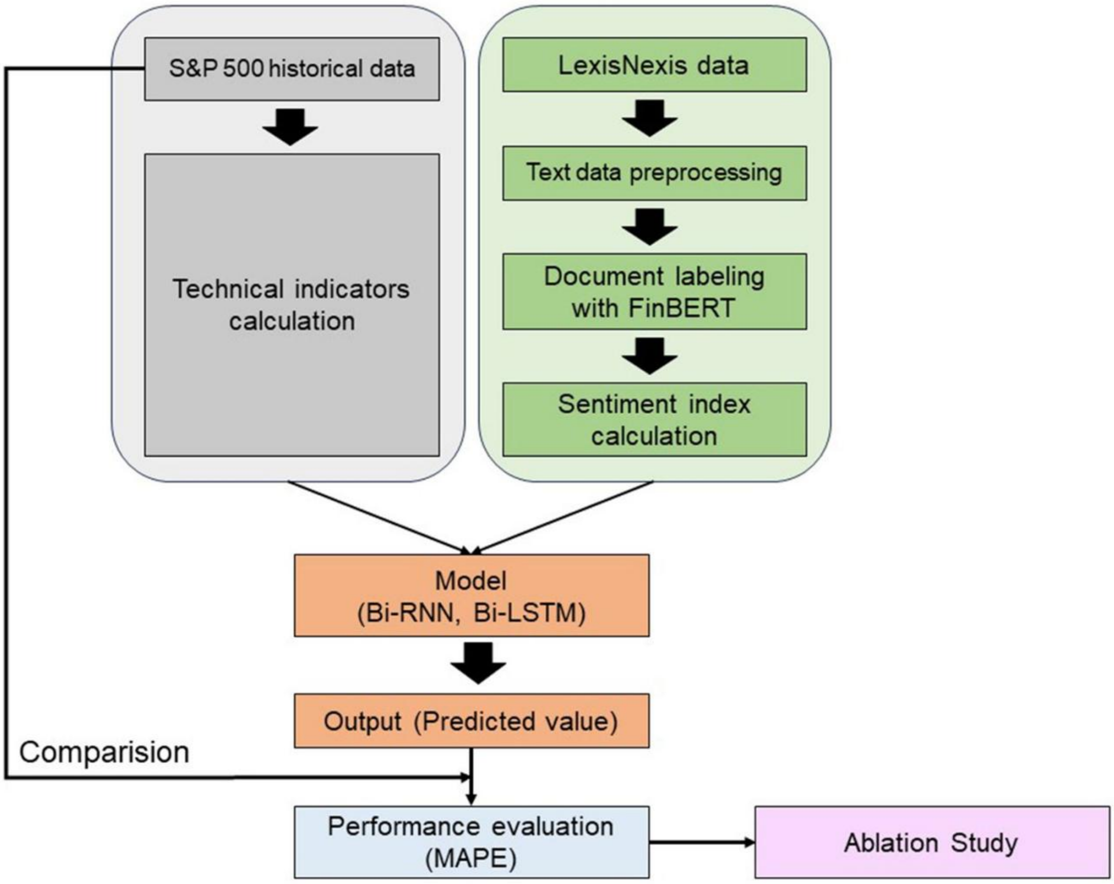

Figure 2. ESG sentiment and technical-indicator prediction pipeline (adapted from [13]).

As shown in Figure 2, historical index data and ESG news text are separated into two distinct processing streams: one dedicated to technical indicator construction and the other focused on sentiment extraction, followed by model-level integration [13]. The content of the image aligns more closely with a pipeline view than with a generic conceptual diagram, effectively illustrating the flow of information and processes involved in this sophisticated analytical approach. This structured methodology highlights the importance of integrating various data types to enhance predictive modeling in financial markets.

A parallel line of development focuses on industry-specific scenarios. One study concentrated on the Chinese pharmaceutical sector and, in addition to stock time series and news text, introduced COVID-related variables while discussing the influence of exogenous shocks on industry trends [25]. This perspective aligns with event-driven stock prediction research, which emphasizes that market prices are influenced not only by endogenous fluctuations but also by various industry events, policy changes, and public risk factors [5]. By incorporating these considerations, researchers aim to enhance the accuracy of predictions and better understand the dynamics that drive stock performance within specific sectors.

As illustrated in Figure 3, reference [25] presents a multi-stock closing-price chart specifically for firms operating within the pharmaceutical sector. Since the current image depicts a trend plot rather than a comprehensive model framework, the accompanying description is thoughtfully aligned with the observed cross-stock heterogeneity in price trajectories during the pandemic period. This context allows for a deeper understanding of how different pharmaceutical companies responded to the market dynamics, reflecting variations that occurred as a result of shifting economic conditions and public health developments. Overall, this analysis sheds light on the intricate relationships between stock performance and external factors affecting the industry during this critical time.

Figure 3. Closing-price trajectories of pharmaceutical stocks during the pandemic period (adapted from [25]).

From a review perspective, the significance of this stage lies in expanding multimodal stock prediction from a text-plus-price setting to a framework that also incorporates long-horizon governance signals and exogenous shocks, thereby improving the interpretability of models under real market conditions.

3.3. Graph Relation Modeling and Stable Fusion Mechanisms

As modality types continue to expand, simple feature concatenation becomes increasingly insufficient for effectively representing the complex dependencies across various information sources. The graph attention network proposed offers a versatile tool for stock relation modeling [24]. Later on, a multi-level financial representation framework was introduced that incorporates macro, sector, and stock perspectives [20]. Additionally, another study further enhanced cross-stock relation modeling with a multiscale multimodal dynamic graph convolution network [15]. These advancements reflect the ongoing efforts to develop more sophisticated methods that can capture intricate relationships within multimodal data, ultimately improving the accuracy of stock predictions.

At the fusion level, the MSGCA framework was proposed, which separately encodes indicator sequences, dynamic documents, and relation graphs before integrating them through a two-stage gated cross-attention mechanism [30]. An important characteristic of this design is that relatively stable indicator modalities act as the guiding branch, facilitating a more robust integration process for the noisier and less balanced text and relation modalities. This thoughtful approach enhances the overall effectiveness of the fusion mechanism, allowing for improved performance in stock prediction tasks by leveraging diverse information sources.

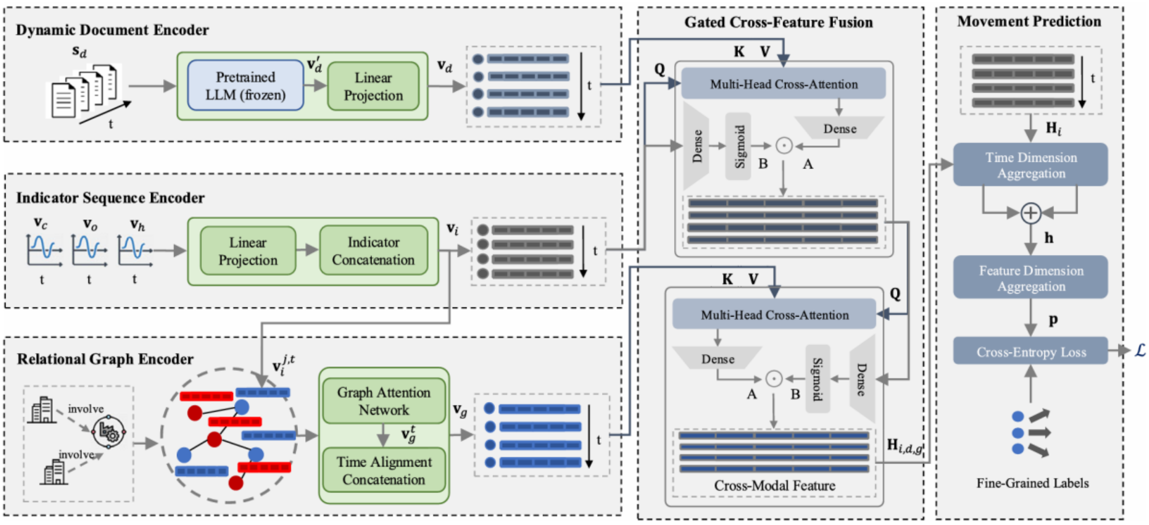

Figure 4. MSGCA framework for stable multimodal fusion (adapted from [30]).

As illustrated in Figure 4, features from indicators, text, and graph relations are progressively integrated using a sophisticated two-stage gated cross-attention mechanism [30]. This approach ensures that the image content is fully aligned with a robust fusion architecture, rather than merely being represented as an outcome plot. By emphasizing this stable integration, the method effectively enhances the coherence and interpretability of the visual data, allowing for more meaningful insights to be derived from the combined features. Overall, this innovative technique demonstrates significant potential for improved analysis through thoughtful feature integration, paving the way for more accurate stock predictions.

The introduction of stable fusion mechanisms marks a significant shift in research emphasis within the field. Rather than merely questioning whether additional modalities should be incorporated, researchers are increasingly focusing on how to effectively manage challenges such as data sparsity, semantic conflicts, and quality disparities among the various modalities. This evolving perspective highlights a deeper understanding of the complexities involved in integrating multimodal data, paving the way for more robust methodologies that can enhance predictive accuracy and overall performance in diverse applications. As such, addressing these issues has become essential for advancing the effectiveness of multimodal approaches in practical scenarios.

3.4. Long-Text Processing, Patch-Based Modeling, and LLM Expansion

A shared characteristic of financial news, corporate announcements, analytical reports, and social commentary is their increasing length and semantic redundancy. To address this issue, a framework was proposed that combines hierarchical LLM-based text processing with a patch-based Transformer, representing one of the most distinctive recent approaches for handling long texts in multimodal stock prediction [28]. This innovative strategy aims to enhance the efficiency and effectiveness of processing lengthy and complex textual data, ultimately improving the accuracy of predictions by better capturing the relevant information embedded within these extended narratives.

As shown in Figure 5, the system is organized into Hierarchical Summary, Embedding and Alignment, Co-Attention Fusion, and Classification stages [28]. The current image corresponds to a comprehensive modeling framework, aligning well with the surrounding methodological discussion. This structured representation underscores the systematic approach taken in integrating various components for effective long-text handling within multimodal stock prediction, facilitating improved analysis and inference from complex data sources.

This line of work is conceptually continuous with Time-LLM [12], zero-shot time-series forecasting with large language models [9], and related surveys on foundation models for time series [11]. In all of these studies, large language models are no longer used solely as tools for text understanding; they also participate in time-series modeling through reprogramming or semantic projection.

Meanwhile, tasks such as FinQA [4], DocFinQA [16], and FinTextQA [22] indicate that financial question answering, report understanding, and numerical reasoning over long documents are becoming important directions in financial intelligence. Combined with PatchTST [19], TimesNet [29], and iTransformer [15], these developments are pushing multimodal stock prediction toward a new stage characterized by semantic compression, structural alignment, and stronger explanatory capacity.

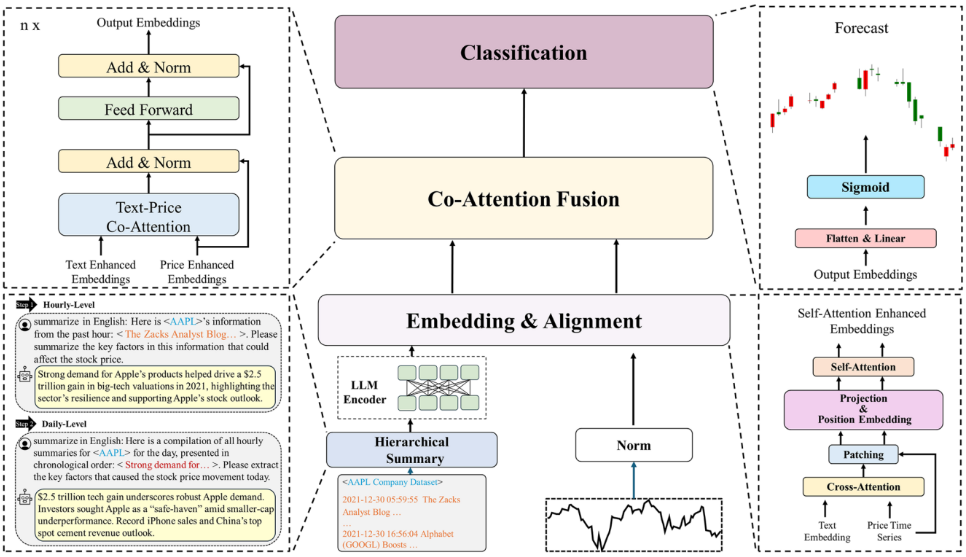

Figure 5. Hierarchical LLM summarization and text-price collaborative modeling framework (adapted from [28]).

4. Major Methodological Categories

Taken together, multimodal stock prediction has developed into a relatively clear and comprehensive technical spectrum. Text-price bimodal fusion serves as the foundational starting point for multimodal modeling, while advancements in ESG factors, industry context, and event enhancement have propelled research toward more complex and realistic market environments. Moreover, the incorporation of graph relations and stable fusion techniques effectively addresses issues related to modality noise and conflicts that may arise during analysis. Additionally, the introduction of Transformers, patch-based modeling approaches, and Large Language Models (LLMs) has significantly expanded the modeling boundaries for both time series data and textual information. This evolution underscores the growing sophistication and adaptability of multimodal methods in capturing the complexities of financial markets.

Table 1. Major methodological categories in multimodal stock prediction.

Method category	Representative studies	Core idea	Main characteristics
Text-price bimodal fusion	[3]; [26]; [21]	Jointly model financial text and historical prices	Establishes the foundation of multimodal prediction
ESG and exogenous enhancement	[13]; [25]; [5]	Introduce ESG, pandemic, policy, and other extended signals	Strengthens interpretability and scenario adaptation
Graph relations and stable fusion	[30]; [24]; [20]; [15]	Use graph structures and gating mechanisms to integrate multi-source information	Addresses modality conflict and relational dependence
Transformer and patch-based temporal modeling	[23]; [19]; [29]; [15]	Replace conventional RNNs with efficient long-sequence modeling	Improves temporal representation capacity
LLMs and long-document processing	[28]; [12]; [9]; [27]; [4]; [16]; [22]	Enhance semantic modeling through summarization, reprogramming, and reasoning	Extends to long-document understanding and richer explanations

5. Key Challenges

5.1. Insufficient Label Granularity and Task Definition

Although multimodal stock prediction studies have continued to expand in both the number of modalities and model complexity, most tasks are still formulated as binary up-down classification or coarse up-flat-down classification. Such label definitions are convenient for standardized evaluation, yet they remain insufficient for representing finer market states such as flat movement, accelerated rise, accelerated fall, and near-reversal conditions.

5.2. Long-Text Redundancy and Insufficient Explicit Event Modeling

Financial news and corporate announcements often contain substantial background information, repetitive narrative, and content that is not directly price sensitive. Although hierarchical LLM summarization [28] and long-document financial tasks such as FinQA [4], DocFinQA [16], and FinTextQA [22] provide new processing routes, comparatively few stock-prediction studies explicitly model event type, impact strength, transmission window, and market feedback path in an integrated manner.

5.3. Modality Conflict, Missing Data, and Fusion Stability

Different modalities are not always consistent with one another. Prices may decline while textual signals remain positive; a highly salient event may receive little immediate price response; and some stocks may suffer from sparse news coverage or weak graph connectivity. Consequently, multimodal fusion is not merely a feature-concatenation problem but also a modality-reliability problem. Works such as MSGCA [30] already recognize this issue, yet stable fusion remains a key challenge that has not been fully resolved.

5.4. Interpretability and Cross-Market Transfer

Financial prediction tasks inherently require strong interpretability. Investors and researchers are concerned not only with whether a prediction is accurate but also with why a model reaches a particular conclusion. However, many current approaches still lack clear explanations of the textual evidence, event drivers, and relation-propagation paths underlying their predictions. At the same time, transferability across markets, sectors, and time periods remains under-discussed.

6. Summary of Research Trends

The existing literature indicates that the main trajectory of multimodal financial stock prediction is becoming increasingly refined.

First, task definitions are moving toward finer-grained state modeling, with growing attention to fine-grained state prediction, reversal identification, and joint risk-return modeling.

Second, explicit event modeling is becoming more closely integrated with long-document understanding, making the interaction among news, announcements, financial reports, and event graphs an important extension of recent research.

Third, stable fusion and interpretability have become more central evaluation dimensions. Reporting accuracy alone is no longer sufficient to reflect the practical value of a model; the ability to explain predictive evidence, recognize modality conflict, and maintain temporal stability is increasingly treated as a marker of research quality.

Fourth, financial large language models and foundation models are significantly propelling the field toward more unified modeling approaches and a deeper understanding of long documents. Models such as FinGPT [27] and Time-LLM [12], along with related studies like [9] and [11], are reshaping the joint modeling of textual information and time series data, fostering a more cohesive analytical framework. Meanwhile, initiatives like FinQA [4], DocFinQA [16], and FinTextQA [22] are accelerating the integration of stock prediction techniques with comprehensive financial document understanding. This convergence not only enhances predictive capabilities but also enriches the contextual insights that can be drawn from complex financial narratives, highlighting the transformative potential of these advanced models in the field of finance.

7. Conclusion

Centered on five recent representative papers and supported by related studies on financial text representation learning, graph relation modeling, patch-based temporal modeling, and financial large language models, this paper has reviewed the research progress of multimodal stock prediction. From an overall perspective, the field has evolved from text-price bimodal fusion to joint modeling over ESG, exogenous events, graph relations, and long-text semantic extraction.

At the same time, insufficient label granularity, long-text redundancy, limited explicit event modeling, unstable multimodal fusion, and insufficient interpretability remain core problems in current multimodal financial prediction. The overall direction visible in the literature shows that finer task definitions, stronger event and relation modeling, and more robust cross-modal fusion mechanisms are becoming the key themes in the continuing evolution of this field.

References

[1] Araci, D. (2019). FinBERT: Financial sentiment analysis with pre-trained language models. *arXiv, arXiv*:1908.10063.

[2] Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. Journal of Computational Science, *2(1), 1-8*.

[3] Chen, P., Boukouvalas, Z., & Corizzo, R. (2024). A deep fusion model for stock market prediction with news headlines and time series data. *Neural Computing and Applications*, 36, 21229-21271.

[4] Chen, Z., Chen, W., Smiley, C., et al. (2021). FinQA: A dataset of numerical reasoning over financial data. *In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 3697-3711).

[5] Ding, X., Zhang, Y., Liu, T., & Duan, J. (2015). Deep learning for event-driven stock prediction. *In Proceedings of the 24th International Joint Conference on Artificial Intelligence (IJCAI)* (pp. 2327-2333).

[6] Dong, Z., Fan, X., & Peng, Z. (2024). FNSPID: A comprehensive financial news dataset in time series. *arXiv, arXiv*:2402.06698.

[7] Feng, F., Chen, H., He, X., Ding, J., Sun, M., & Chua, T.-S. (2019). Enhancing stock movement prediction with adversarial training. *In Proceedings of the 28th International Joint Conference on Artificial Intelligence (IJCAI)* (pp. 5843-5849).

[8] Fischer, T., & Krauss, C. (2018). Deep learning with long short-term memory networks for financial market predictions. *European Journal of Operational Research*, *270*(2), 654-669.

[9] Gruver, N., Finzi, M., Qehaja, M., & Wilson, A. G. (2023). Large language models are zero-shot time series forecasters. *In Advances in Neural Information Processing Systems*.

[10] Gu, A., & Dao, T. (2023). Mamba: Linear-time sequence modeling with selective state spaces. *arXiv, arXiv*:2312.00752.

[11] Jin, M., Wang, S., Ma, L., Chu, Z., Zhang, J. Y., & Wen, Q. (2023). Time series forecasting with pretrained foundation models: A survey. *arXiv, arXiv*:2312.17119.

[12] Hu, Z., Zhao, Y., & Khushi, M. (2023). Time-LLM: Time series forecasting by reprogramming large language models. *arXiv, arXiv*:2310.01728.

[13] Lee, H., Kim, J. H., & Jung, H. S. (2024). Deep-learning-based stock market prediction incorporating ESG sentiment and technical indicators. *Scientific Reports*, 14, Article 10262.

[14] Liu, Y., Hu, T., Zhang, H., Wu, H., Wang, S., Ma, L., & Long, M. (2024). iTransformer: Inverted transformers are effective for time series forecasting. *In Proceedings of the International Conference on Learning Representations (ICLR).*

[15] Liu, Z., Chen, C., Wang, B., et al. (2024). Melody-GCN: A multiscale multimodal dynamic graph convolution network for stock forecasting. *Pattern Recognition*, 61(1).

[16] Lu, Y., Xing, H., & Zhou, C. (2024). DocFinQA: A long-context financial reasoning dataset. *In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*.

[17] Maia, M., Handschuh, S., Freitas, A., et al. (2018). WWW'18 open challenge: Financial opinion mining and question answering. *In Companion Proceedings of the Web Conference 2018*.

[18] Malo, P., Sinha, A., Korhonen, P., Wallenius, J., & Takala, P. (2014). Good debt or bad debt: Detecting semantic orientations in economic texts. *Journal of the Association for Information Science and Technology*, 65(4), 782-796.

[19] Nie, Y., Nguyen, N. H., Sinthong, P., & Kalagnanam, J. (2022). A time series is worth 64 words: Long-term forecasting with transformers. *arXiv, arXiv*:2211.14730.

[20] Qin, W., Song, Y., Chen, W., et al. (2021). Coupling macro-sector-micro financial indicators for learning stock representations with less uncertainty. *In Proceedings of the 35th AAAI Conference on Artificial Intelligence* (pp. 9122-9129).

[21] Sawhney, R., Agarwal, S., Wadhwa, A., & Shah, R. R. (2020). Deep attentive learning for stock movement prediction from social media text and company correlations. *In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (pp. 8415-8426).

[22] Shen, T., Zhang, S., Li, Y., et al. (2024). FinTextQA: A dataset for long-form financial question answering. *In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (ACL)*.

[23] Vaswani, A., Shazeer, N., Parmar, N., et al. (2017). Attention is all you need. *In Advances in Neural Information Processing Systems (Vol. 30)*.

[24] Velickovic, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., & Bengio, Y. (2017). Graph attention networks. *arXiv, arXiv*:1710.10903.

[25] Wang, H., Xie, Z., Chiu, D. K. W., & Ho, K. K. W. (2025). Multimodal market information fusion for stock price trend prediction in the pharmaceutical sector. *Applied Intelligence*, 55, Article 77.

[26] Xu, Y., & Cohen, S. B. (2018). Stock movement prediction from tweets and historical prices. *In Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL)* (pp. 1970-1979).

[27] Yang, Y., Xie, Y., Wang, C., et al. (2023). FinGPT: Open-source financial large language models. *arXiv, arXiv*:2306.06031.

[28] Zhang, Y., Dong, Z., & Xu, W. (2026). Integrative stock price trend prediction via hierarchical LLM text processing and patch-based transformer with co-attention. *Expert Systems with Applications*, 302, Article 130441.

[29] Zhou, T., Zhang, P., & Yan, Z. (2023). TimesNet: Temporal 2D-variation modeling for general time series analysis. *In Proceedings of the International Conference on Learning Representations (ICLR)*.

[30] Zong, C., Wan, J., Cascone, L., & Zhou, H. (2025). Stock movement prediction with multimodal stable fusion via gated cross-attention mechanism. *Complex & Intelligent Systems*, 11, Article 396.