

Multimodal Market-Risk Forecasting: A Systematic Review of Fusion, Generalisability, and Deployment Readiness

Yihao Ren

*School of Computer Science and Technology, Beijing Jiaotong University, Beijing, China
23722023@bjtu.edu.cn*

Abstract. Multimodal models increasingly combine market series with text, audio, macroeconomic indicators, and relational data, but their generalisability and deployment readiness remain uncertain. We systematically reviewed studies published from 1 January 2015 to 20 July 2026 and complemented the review with keyword co-occurrence analysis, verified quantitative direction, quality appraisal, and a three-dimensional evidence-support evaluation. One hundred studies met the eligibility criteria; 68 formally indexed English-language records formed the bibliometric corpus and 67 were retained by CiteSpace. Volatility forecasting accounted for 77 studies and equity applications for 63. Feature/covariate and graph/relational integration together represented 68 studies. Among 45 studies with page-verified quantitative evidence, 33 contained at least one favourable focal-model comparison, although heterogeneous targets, horizons, metrics, and baselines precluded meta-analysis. No study demonstrated independent external validation under the strict criterion. However, 58 reported temporal generalisation and 22 cross-dataset or cross-domain replication; 66 met at least one partial criterion. Timeliness or deployment latency was unreported in 91 studies, and audit, governance, or model-monitoring evidence was not located. The evidence therefore supports scenario-specific method selection, but not a universal architecture ranking or a comparative accuracy-latency-cost frontier. Progress depends on independent testing, timestamp transparency, reproducible pipelines, and end-to-end operational evaluation.

Keywords: multimodal learning, financial market risk, data fusion, generalisability, model governance

1. Introduction

Market risk is rarely contained in a single data stream. Prices and realised measures encode trading outcomes, while macroeconomic releases, news, social media, corporate disclosures, and inter-market relations provide information about expectations and transmission mechanisms. This makes volatility, value-at-risk (VaR), tail-risk, and extreme-event forecasting natural candidates for multimodal learning. Early studies combined prices with news or sentiment indices; later work introduced verbal-acoustic fusion, graph representations, joint encoders, and large language models [1-5].

The expansion of model capacity has not resolved a more basic evidential problem. A favourable result on a single test set does not show that a model transfers across market regimes or institutions, and a low model-inference time does not demonstrate that data collection, preprocessing, and updating can operate within a decision deadline. Similarly, the presence of attention weights or feature-importance plots does not establish auditability, stability, or regulatory compliance. These distinctions are especially important in finance, where information availability is time-dependent and deployment decisions may affect capital allocation or risk limits.

Existing reviews address financial text analysis, deep learning for prices, or sentiment forecasting, but rarely connect fusion design to temporal validity, transportability, latency, and governance. A simple comparison of reported accuracy is also problematic because studies differ in market, horizon, target construction, baseline strength, and metric direction.

This review addresses five questions. First, how has the scale and knowledge structure of multimodal market-risk forecasting evolved since 2015? Second, which modalities, fusion architectures, markets, risk targets, datasets, and metrics have been used? Third, how do studies handle trade-offs among predictive performance, data timeliness, inference latency, and deployment cost? Fourth, how do approaches differ in technical feasibility, scenario fit, and regulatory readiness? Fifth, which conclusions are supported by stable evidence, and which remain constrained by temporal leakage, limited reproducibility, missing latency reporting, or sparse governance evidence?

We make four contributions. We construct a market-risk-specific corpus rather than treating generic price-direction prediction as risk forecasting. We develop a hierarchical classification that separates observable information integration from more deeply learned cross-modal interaction. We distinguish strict independent external validation from partial temporal and cross-domain generalisation. Finally, we integrate bibliometric structure, verified quantitative direction, quality appraisal, and a three-dimensional evidence-support framework. The objective is not to nominate a universally best model, but to identify where the evidence is sufficiently mature to guide method choice and where it is not.

2. Methods

2.1. Review design and reporting

The study combined a systematic review, structured narrative synthesis, keyword co-occurrence analysis, and evidence-support evaluation. Reporting followed PRISMA 2020 where applicable [6].

2.2. Information sources and search strategy

The core search covered Web of Science Core Collection, Scopus, IEEE Xplore, and the ACM Digital Library. arXiv and SSRN were used to identify frontier work. Search concepts covered four blocks: multimodality or data fusion; financial markets; volatility, VaR, tail risk, or extreme events; and forecasting or early warning. Searches covered records published from 1 January 2015 to 20 July 2026; the final formal search was completed on 20 July 2026.

Records were deduplicated first by normalised DOI and then by normalised title, authorship, and year. The principal English sources yielded 2,770 records, of which 100 studies were included after screening (Figure 1).

PRISMA 2020 study-selection flow

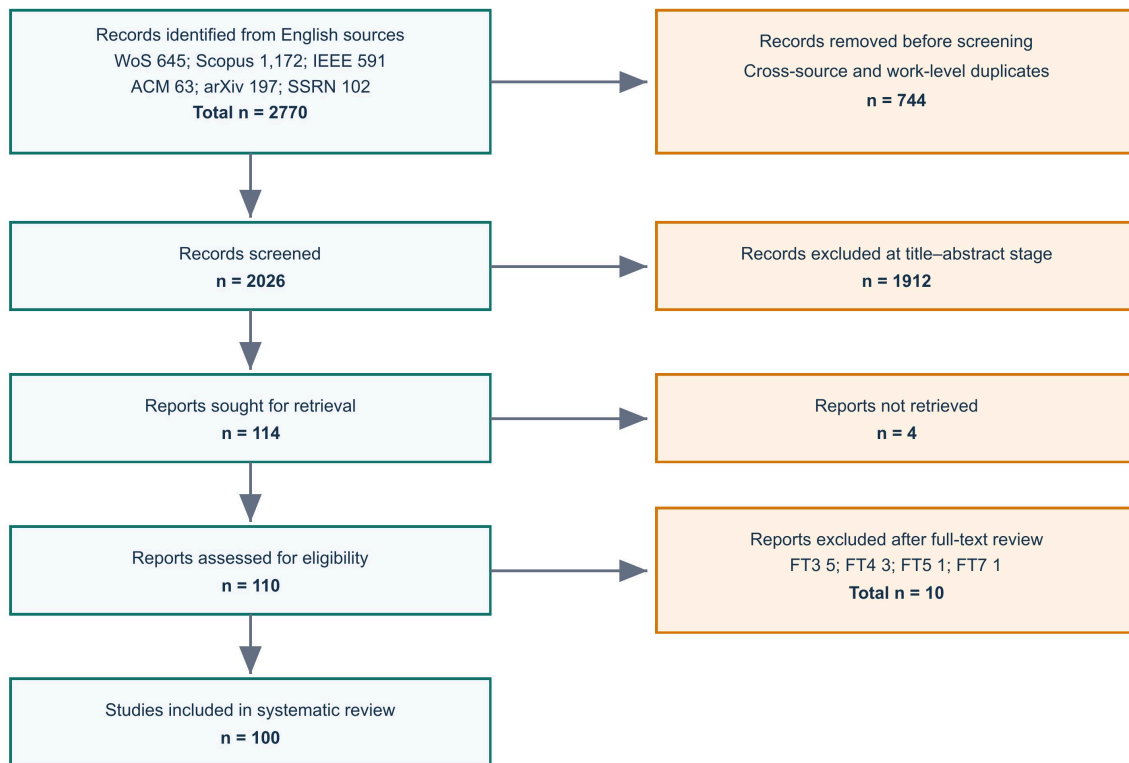


Figure 1. PRISMA 2020 flow diagram for the English core review

2.3. Eligibility criteria

Eligible studies combined at least two substantively different information modalities or sources and evaluated an original predictive model for an explicit market-risk target. Qualifying targets included volatility, VaR or quantile risk, tail or downside risk, price-crash risk, and defined extreme-event warnings. Studies restricted to ordinary return, price, or direction prediction without a risk definition were excluded, as were credit risk, fraud, anti-money-laundering, insurance, pure asset allocation, and non-empirical summaries. Peer-reviewed work formed the core evidence; relevant preprints were retained as a separately labelled frontier layer.

2.4. Data extraction and hierarchical classification

For each study, we recorded publication status, market, risk task, modalities, fusion method, sampling frequency, forecast horizon, data-availability timing, train-validation-test design, models, baselines, ablations, outcome metrics, reproducibility evidence, timeliness evidence, and regulatory-readiness indicators.

The hierarchical taxonomy classified the observable integration pathway before assigning a more specific subtype. The main families were feature/covariate integration, graph/relational integration, prediction or decision combination, joint representation learning, cross-modal interaction, large-language-model-mediated semantic integration, and multi-source integration whose pathway was

not clearly reported. This approach avoided inferring a mechanism from an abstract or model name. Only three studies remained at the broad multi-source level after full-text recoding.

2.5. Quality, generalisation, and timeliness appraisal

Quality appraisal covered target validity, temporal integrity, evaluation completeness, baseline adequacy, robustness, external validity, and reproducibility. These domains assessed whether the risk outcome was well defined, evaluation respected information timing, comparisons and sensitivity checks were adequate, findings extended beyond the development setting, and the analysis could be independently reconstructed.

Strict external validation (EV_STRICT) required evaluation on a genuinely independent population, market, institution, geography, or independently collected dataset outside model development and tuning. A separate generalisation layer captured basic chronological or out-of-sample evaluation (T1), enhanced rolling, recursive, repeated, or regime-aware temporal evaluation (T2), cross-dataset or cross-domain replication (D1), source-to-target transfer without full target redevelopment (D2), and cross-market or cross-asset replication (M1). T2 was a subset of T1, while D2 and M1 were subsets of D1. Partial criteria were not relabelled as external validation. Studies meeting one or more of these criteria were classified as providing partial generalisation evidence.

Timeliness was graded from no qualifying timing evidence (NR) through model-level timing to end-to-end or live evidence. Statements about future deployment or computational efficiency without measured timing did not receive a positive latency code.

2.6. Quantitative and narrative synthesis

Studies with page-verified numerical comparisons and sufficiently interpretable experimental settings formed the Tier-A quantitative layer. We preserved the original metric, unit, direction, baseline, horizon, and dataset. Focal-model comparisons were classified as favourable, unfavourable, equal, or indeterminate. Because metrics, horizons, markets, and comparators were not exchangeable, no pooled effect size was calculated. Study-level direction indicates whether at least one recorded comparison favoured the focal approach; it does not isolate multimodality as the causal source of improvement.

2.7. Bibliometric analysis and three-dimensional evidence-support evaluation

We conducted keyword co-occurrence analysis using CiteSpace [7] on 68 formally indexed English-language records, with a prespecified g-index baseline of $k=25$ and sensitivity tests at $k=15$ and $k=50$.

Each study was scored on technical feasibility, scenario fit, and regulatory readiness. Scores of 0, 1, and 2 indicated no demonstrated support, partial support, and explicit support, respectively. Missingness was reported separately. Equal weights formed the baseline; technical-, business-, and regulatory-priority scenarios and available-case normalisation tested sensitivity.

3. Results

3.1. Corpus profile

Table 1. Characteristics of the included evidence base

Characteristic	Category	n/N	%
Risk task	Volatility forecasting	77/100	77.0
Market	Equity/stock market	63/100	63.0
Publication period	2025-2026	46/100	46.0
Integration	Feature/covariate + graph/relational	68/100	68.0
Integration	Joint/cross-modal/LLM-mediated	21/100	21.0
Fusion reporting	Pathway not clearly reported	3/100	3.0
Quantitative layer	Tier-A with verified numeric rows	45/100	45.0
Bibliometric layer	Formally indexed English records	68/100	68.0

The corpus was concentrated in volatility forecasting (77/100) and equity or stock-market applications (63/100). Forty-six studies were published in 2025-2026, demonstrating that the evidence base is recent, although this proportion is not an annual growth rate and 2026 coverage is incomplete. Other applications included commodities, cryptocurrencies, global indices, VaR, tail risk, and crash warnings.

Feature/covariate and graph/relational integration together accounted for 68/100 studies. Joint representation, cross-modal interaction, and LLM-mediated semantic integration represented 21/100. The distribution suggests that the field has expanded beyond simple feature concatenation, but most studies still organise heterogeneity through engineered predictors or explicit relations rather than fully learned cross-modal interaction.

3.2. Knowledge structure

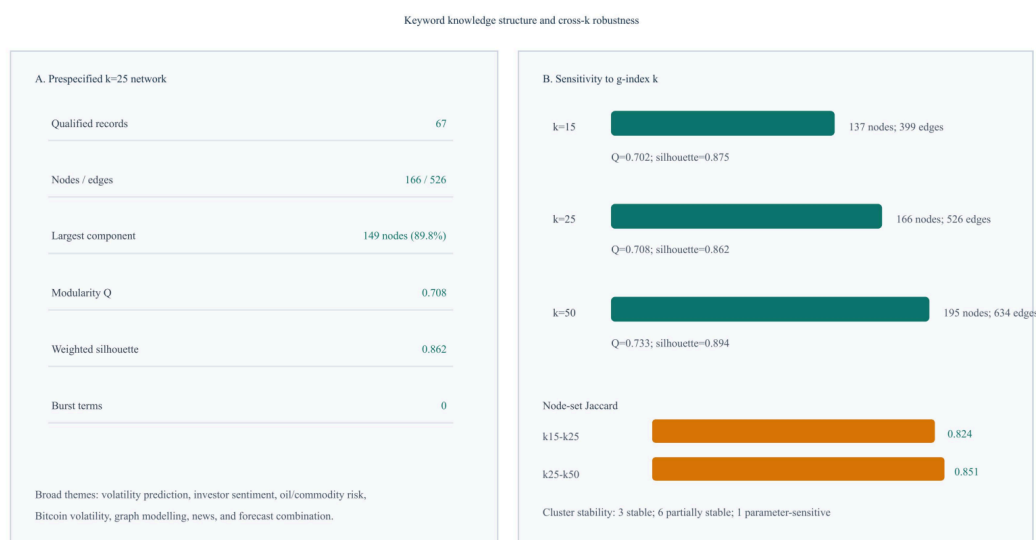


Figure 2. Keyword-network structure and g-index sensitivity

The keyword co-occurrence network ($k=25$) showed a well-formed structure with high modularity ($Q=0.708$) and a weighted mean silhouette coefficient of 0.862 for the principal clusters. Node membership was robust to g -index variation (the $k=15$ and $k=25$ sets had a Jaccard overlap of 0.824, while $k=25$ and $k=50$ had an overlap of 0.851), while cluster boundaries were moderately parameter-sensitive. No burst terms were detected, indicating a stable knowledge core rather than rapidly shifting specialities.

3.3. Modalities, fusion mechanisms, and evidence-aligned applicability

Numerical market series were commonly combined with news, social media, macroeconomic variables, corporate disclosures, audio, order-book information, or relational graphs. Representative work illustrates the progression from news-augmented and social-media forecasting to hierarchical text-price models, verbal-acoustic fusion, graph convolution, alternative-data systems, and LLM-based multi-source systems [1, 3, 5, 8-10].

The evidence supports conditional method selection rather than an architecture ranking. Feature/covariate integration has the widest evidence base and is most directly aligned with daily or lower-frequency settings in which structured predictors and interpretable baselines are important. Graph/relational integration is especially aligned with cross-asset and cross-market settings because it represents dependence explicitly; five of seven cross-market/global-index studies used this family. This concentration is consistent with the relational inductive bias of graph models, although it does not establish superior accuracy.

Audio-enhanced and cross-modal systems address event-driven settings such as earnings calls, where tone and delivery may complement transcript content [3, 11, 12]. Their applicability is narrower because speech is available only around particular events and evidence is based on fewer studies. LLM-mediated integration can process long and heterogeneous text sources, but current evidence does not establish its latency, cost, or stability under operational use. These conclusions are summarised in Table 2.

Table 2. Evidence-aligned method-scenario applicability and trade-off matrix

Method family	Evidence-aligned use	Observed strength	Unresolved condition
Feature/covariate	Daily or lower-frequency risk with structured predictors and auditable baselines	Widest evidence base; favourable comparisons are frequent	Marginal value depends on feature design; latency and cost rarely reported
Graph/relational	Cross-asset, cross-market, spillover, and connected-risk settings	Explicitly represents dependence; frequent in cross-market studies	Graph construction, updating, and operational burden were not compared
Joint/cross-modal	Tasks requiring nonlinear interaction across text, audio, and market series	Can represent richer interaction than static concatenation	Temporal integrity, ablations, and transportability remain uneven
Audio-enhanced	Earnings-call and event-driven volatility or risk	Provides a distinct non-verbal information channel	Small evidence base and event-limited data availability
LLM-mediated	Long, heterogeneous, semantic information sources	Flexible multi-source extraction and representation	End-to-end latency, cost, privacy, and stability are under-evidenced

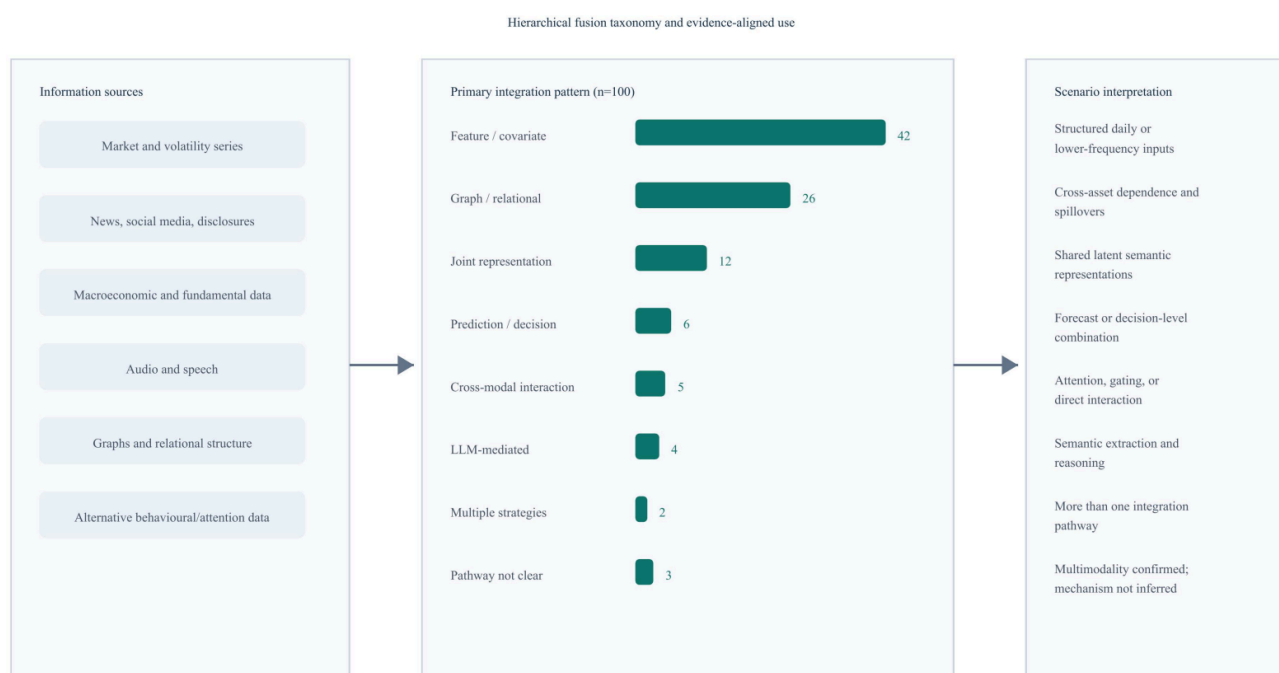


Figure 3. Hierarchical taxonomy linking information sources, primary integration patterns, and evidence-aligned scenario interpretation

3.4. Direction of quantitative evidence

Forty-five Tier-A studies generated 130 verified quantitative rows. Thirty-three studies contained at least one favourable focal-model comparison. Five of these also contained an unfavourable comparison, ten were directionally indeterminate, and two were unfavourable-only.

Thus, favourable results were common but not universal. Several well-specified studies showed gains from news, sentiment, graph structure, or transformer representations [13-15], while others produced mixed or unfavourable results. The most defensible synthesis is that multimodal methods frequently improved at least one reported comparison within their own experimental settings. The evidence does not support a single pooled benefit or the causal assertion that adding modalities is sufficient for improvement.

Among studies with verified direction, 7/10 graph studies and 2/4 audio studies contained a favourable comparison (Fisher exact $p=0.580$). Thus, graph/relational and audio-enhanced approaches did not show a systematic difference in quantitative direction.

3.5. Quality and temporal integrity

Overall appraisal classified 95 studies as having some concerns, three as high concern, and two as low concern; this reflected a stringent multi-domain assessment rather than study invalidity.

Temporal integrity was low risk in 58/100, unclear in 38/100, high risk in 3/100, and of some concern in 1/100. Uncertainty mainly arose from incomplete reporting of time splits and information availability, not confirmed leakage; notably, four of five cross-modal-interaction studies were unclear. Although complex pipelines may encourage greater emphasis on architecture than timestamp documentation, this explanation was not directly tested.

The 13 studies using five or more modalities were not systematically lower in quality than the 78 using three or four. They reported code more often but provided no qualifying timeliness evidence,

leaving their operational burden unresolved.

Quality trends across 2015–2020, 2021–2023, and 2024–2026 were non-monotonic: recent studies showed slightly greater attention to timeliness but no broad improvement in reproducibility or overall quality.

3.6. From temporal testing to external generalisability

No study met the strict criterion for independent external validation (0/100), although partial generalisability evidence was common. Temporal generalisation (T1) appeared in 58/100 studies, including 30/100 with enhanced designs (T2), whereas cross-dataset or cross-domain replication (D1) appeared in 22/100, including 18/100 with cross-market or cross-asset replication (M1); only one study qualified for source-to-target transfer (D2). Overall, 66/100 met T1 or D1, 14/100 met both, and 34/100 met neither. Temporal evidence was more prevalent than cross-domain replication (58% vs. 22%; difference 36 percentage points; exact McNemar $p < 0.001$).

The pattern indicates a broad but incomplete effort to move beyond in-sample performance. One plausible reason for the imbalance is practical: a chronological holdout can be constructed within an existing dataset, whereas independent replication requires another market, institution, or data-generation process and may involve licensing, harmonisation, and governance constraints. But a time-ordered test within the same pipeline cannot establish transportability. The field is therefore closer to routine temporal evaluation than to independent external validity.

3.7. Accuracy, timeliness, latency, and deployment cost

The literature largely assessed predictive performance in isolation: 91/100 studies reported no qualifying timeliness or deployment-latency evidence, six reported model-level timing (RT1), two only limited computational evidence (RT0), and one end-to-end or live evidence (RT2). Hardware, energy, data-acquisition, and updating costs were not comparable across studies. The evidence therefore cannot determine whether accuracy gains offset latency, computational demand, or deployment expense, and no comparable accuracy-latency-cost frontier has been established. This gap may reflect offline benchmarking conventions that prioritise predictive metrics over pipeline timing and operating cost.

3.8. Reproducibility and regulatory readiness

Seventy-one studies had a high reproducibility risk, indicating limited support for independent reconstruction. explicit code and data availability were found in 21/100 and 13/100, respectively. Interpretability and robustness evidence appeared in 45/100 and 56/100, while fairness/bias and privacy appeared in 10/100 each; no audit, governance, or model-monitoring evidence was located.

The three-dimensional scores were consistent with these patterns. Mean technical, scenario, and regulatory evidence-support scores were 32.4, 56.6, and 24.2, with criterion missingness of 39.0%, 34.0%, and 75.8%, respectively. Ranking correlations with the equal-weight baseline were high under technical-, business-, and regulatory-priority weights, but available-case normalisation reduced the correlation to 0.756 and the top-ten overlap to three. Missing evidence therefore materially affects study-level rankings.

Application-scenario fit was documented more often than operational or governance readiness. The low regulatory score may reflect experimental reporting rather than non-compliance, but absent documentation cannot support a readiness claim.



Figure 4. Three-dimensional evidence support, generalisation breadth, and timeliness evidence

4. Discussion

4.1. What multimodality currently contributes

Multimodal market-risk research has developed along three main paths. The first expands the predictor set by adding sentiment, macroeconomic, search, disclosure, or order-book variables to a time-series model. The second makes relations among assets or markets explicit through graphs. The third learns joint semantic representations across text, audio, market data, and, increasingly, LLM-derived features. These approaches address different information problems, so more complex fusion should not be assumed to be universally more advanced or more suitable.

Feature/covariate integration remains useful because it can preserve strong econometric or machine-learning baselines and keep the contribution of each source relatively visible. Graph methods are better aligned with questions in which connectedness is part of the target mechanism, such as spillovers or cross-market volatility [14, 16]. Audio and deep cross-modal models are most defensible when non-verbal or interaction effects are central to the information event. LLM systems broaden semantic coverage but currently add an unmeasured operational layer.

4.2. Predictive gains versus evidential strength

Favourable comparisons were common: 33/45 quantitatively verified studies reported at least one, suggesting that multimodal information can add predictive value within a study's chosen data, horizon, baseline, and metric. However, this does not show that adding modalities or increasing fusion complexity is universally beneficial. Statistical significance or uncertainty was reported in only 5/45 studies, and no study met the strict external-validation criterion (0/100), limiting claims about robustness and transportability. Partial generalisation evidence in 66/100 studies places the field at an intermediate stage: temporal evaluation is relatively common, but independent transportability remains rare. Because 91/100 studies lacked qualifying timing evidence, no

comparative accuracy-latency-cost frontier could be established. Operational real time requires timely inputs and reliable end-to-end updating, not fast inference alone.

4.3. Reproducibility, governance, and future priorities

Support for independent reconstruction was limited: 71/100 studies were rated at high reproducibility risk, and no qualifying audit, governance, or model-monitoring evidence was located. This represents an evidence and reporting gap rather than a legal judgement, but no method family can be considered deployment-ready from the present literature alone. Future research should prioritise timestamp audits, locked external tests, cross-market benchmarks, modality-missingness stress tests, and paired reporting of performance, end-to-end latency, and resource demand. Ablations should distinguish gains from new information, fusion architecture, pretraining, and model capacity, while deployment-oriented studies should document data lineage, drift monitoring, escalation thresholds, and retraining controls.

4.4. Limitations

The review was limited to the specified databases, dates, languages, and market-risk criteria. Eligibility screening and evidence synthesis were restricted to English-language studies, so contributions in other languages may therefore be underrepresented. Coverage of 2026 was incomplete, and the bibliometric analysis used a smaller indexed subset of the systematic corpus. Heterogeneity precluded meta-analysis, while quality judgements reflected reported evidence rather than independent replication. The applicability matrix combines observed patterns with explicitly labelled interpretation and is not a causal estimate of method suitability.

5. Conclusion

Multimodal learning has become an established approach to financial market-risk forecasting, particularly for volatility and equity applications. The literature frequently reports favourable predictive comparisons and has developed identifiable method families for structured predictors, relational markets, event-based speech, and heterogeneous semantic information. These families are useful in different settings; no universal ranking is supported.

The field has also made meaningful, though incomplete, progress towards generalisation. Two-thirds of the studies used at least one temporal or cross-domain evaluation strategy, but none demonstrated independent external validation under the strict criterion. Evidence of end-to-end latency, deployment cost, reproducibility, and governance was considerably weaker than evidence of task and horizon alignment.

The clearest practical conclusion is therefore conditional. Feature-based systems currently offer the broadest evidence base for structured forecasting; graph models are well aligned with relational and cross-market risk; audio and deep semantic fusion address narrower information-rich events; and LLM-mediated systems remain promising but operationally under-evidenced. Progress now depends less on adding modalities indiscriminately than on demonstrating that predictive gains survive time, market change, operational constraints, and governed use.

References

- [1] Atkins, A., Niranjana, M., & Gerding, E. (2018). Financial news predicts stock market volatility better than close price. *Journal of Finance and Data Science*. <https://doi.org/10.1016/j.jfds.2018.02.002>

- [2] Qin, Y., & Yang, Y. (2019). What you say and how you say it matters: Predicting stock volatility using verbal and vocal cues. *Annual Meeting of the Association for Computational Linguistics*. <https://doi.org/10.18653/v1/P19-1038>
- [3] Sawhney, R., Khanna, P., Aggarwal, A., Jain, T., Mathur, P., & Shah, R. R. (2020). VolTAGE: Volatility forecasting via text-audio fusion with graph convolution networks for earnings calls. *Conference on Empirical Methods in Natural Language Processing*. <https://doi.org/10.18653/v1/2020.emnlp-main.643>
- [4] Sawhney, R., Mathur, P., Mangal, A., Khanna, P., Shah, R. R., & Zimmermann, R. (2020). Multimodal multi-task financial risk forecasting. *MM 2020 - Proceedings of the 28th ACM International Conference on Multimedia*. <https://doi.org/10.1145/3394171.3413752>
- [5] Cao, Y., Chen, Z., Kumar, P., Pei, Q., Yu, Y., Li, H., Dimino, F., Ausiello, L., Subbalakshmi, K., Ndiaye, P. M., & Li, H. (2024). RiskLabs: Predicting financial risk using large language model based on multimodal and multi-sources data. *2025 IEEE International Conference on Data Mining Workshops (ICDMW)*, 901-909. <https://doi.org/10.1109/icdmw69685.2025.00107>
- [6] Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., Shamseer, L., Tetzlaff, J. M., Akl, E. A., Brennan, S. E., Chou, R., Glanville, J., Grimshaw, J. M., Hróbjartsson, A., Lalu, M. M., Li, T., Loder, E. W., Mayo-Wilson, E., McDonald, S., ... Moher, D. (2021). The PRISMA 2020 statement: An updated guideline for reporting systematic reviews. *BMJ*, 372, n71. <https://doi.org/10.1136/bmj.n71>
- [7] Chen, C. (2006). CiteSpace II: Detecting and visualizing emerging trends and transient patterns in scientific literature. *Journal of the American Society for Information Science and Technology*, 57(3), 359-377. <https://doi.org/10.1002/asi.20317>
- [8] Yang, L., Ng, T. L. J., Smyth, B., & Dong, R. (2020). HTML: Hierarchical transformer-based multi-task learning for volatility prediction. *Proceedings of The Web Conference 2020*. <https://doi.org/10.1145/3366423.3380128>
- [9] Akbiyik, M. E., Erkul, M., Kämpf, K., Vasiliauskaite, V., & Antulov-Fantulin, N. (2023). Ask "who", not "what": Bitcoin volatility forecasting with Twitter data. *WSDM 2023 - Proceedings of the 16th ACM International Conference on Web Search and Data Mining*. <https://doi.org/10.1145/3539597.3570387>
- [10] Rahimikia, E., & Poon, S. (2026). Alternative data for realised volatility forecasting: Limit order book and news stories. *International Journal of Finance & Economics*. <https://doi.org/10.1002/ijfe.70239>
- [11] Sawhney, R., Aggarwal, A., & Shah, R. R. (2021). An empirical investigation of bias in the multimodal analysis of financial earnings calls. *NAACL-HLT 2021 - 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Proceedings of the Conference*. <https://doi.org/10.18653/v1/2021.naacl-main.294>
- [12] Wang, S., Ji, T., He, J., Almutairi, M., Wang, D., Wang, L., Zhang, M., & Lu, C.-T. (2024). AMA-LSTM: Pioneering robust and fair financial audio analysis for stock volatility prediction. *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL 2024*. <https://doi.org/10.18653/v1/2024.naacl-industry.32>
- [13] Bodilsen, S. T., & Lunde, A. (2025). Exploiting news analytics for volatility forecasting. *Journal of Applied Econometrics*. <https://doi.org/10.1002/jae.3095>
- [14] Gong, J., Wang, G.-J., Zhou, Y., & Xie, C. (2025). Cross-market volatility forecasting with attention-based spatial-temporal graph convolutional networks. *Journal of Empirical Finance*. <https://doi.org/10.1016/j.jempfin.2025.101639>
- [15] Song, Y., Zhang, Y., Huang, J., & Yang, A. (2025). Volatility and value-at-risk forecasting using BERT and transformer models incorporating investors' textual sentiments. *Finance Research Letters*. <https://doi.org/10.1016/j.frl.2025.108210>
- [16] Chi, Z., Gao, J., & Wang, C. (2024). Graph signal processing for global stock market realized volatility forecasting [Preprint]. *arXiv*. <https://arxiv.org/abs/2410.22706>