International Journal of Computer Information Systems and Industrial Management Applications
ISSN 2150-7988 Volume 18(2) 2026 pp. 451-459
https://doi.org/10.70917/ijcisim-2026-2807
Article

Multimodal Representation Learning for High-Stakes Investment Decisions: A Systematic Review and the Multimodal Investment Readiness (MIR) Framework

Shivani Chaudhary, Kanishka Dev Chourasia, Suman Gugulothu
chaudha.shivi@gmail.com; kanishkdev01@gmail.com; sumangugug@gmail.com

Abstract: Investment decisions in institutional settings increasingly depend on data that come from different sources and are not always easy to combine reliably. In this paper, we use the MIR framework to evaluate multimodal representation learning methods, with a focus on how well they translate from model development into practical institutional use. More and more, financial decision-making systems use different types of data, such as market time series, text, audio-derived signals, satellite images, and graph-structured linkages. A common problem with many old methods is that they analyze various inputs individually. This makes it hard to see how signals interact across modalities and can leave the system open to attack if one source is absent or inaccurate. Multimodal representation learning provides an effective alternative by acquiring common representations across diverse data sources instead of analyzing each input independently. This study analyzes the current research from five pragmatic viewpoints: the methodologies employed in finance, the predominant combinatorial strategies identified in prior studies, the actual outcomes in financial applications, the challenges of implementing these methodologies within institutions, and the approaches to evaluate these systems. Based on this research, I propose the MIR framework. MIR includes three useful parts that are typically missing from other projects. The first is modality trust scoring, which lets the model change how much weight it gives to each input based on how reliable it is in a certain situation. The second kind is event-focused alignment contracts. These contracts spell out how modality pairings should match up across time and can also help with model-risk controls. The third kind is uncertainty-to-action layers. These layers link model uncertainty to actions that affect the portfolio, such as sizing, hedging, or deferring execution. The framework is meant to bring the goals of machine learning research and the needs of institutional investment closer together. In institutional investing, governance, auditability, and risk oversight are just as essential as predictive performance.

Keywords: multimodal learning, representation learning, financial machine learning, earnings conference calls, alternative data, uncertainty quantification, model risk management, investment readiness

1. Introduction

Investment decision-making systems presently integrate data from a multitude of sources, a scale unprecedented in its scope. The alternative data landscape has expanded to encompass over 400 specialized providers, and a significant 53% of surveyed hedge fund managers now incorporate alternative data into their investment strategies (AIMA, 2020). Contemporary market signals encompass structured microstructure data, regulatory filings, real-time news, audio from earnings calls, satellite imagery, and relational network architectures. Traditional quantitative methodologies, which typically involve the individual engineering of characteristics from each source and their subsequent aggregation through linear scoring or basic ensembles, frequently fail to account for cross-modal interactions and demonstrate vulnerability when a modality fails (Baltrusaitis et al., 2019).

Multimodal representation learning is an effective way to model information from different sources, especially when these sources share signals rather than functioning independently. Previous studies have demonstrated that shared representations of different information formats can enhance performance in tasks where no single source offers a complete perspective. This idea is starting to be used in finance. Research has indicated improvements in volatility prediction by integrating earnings-call text with audio, return prediction with satellite imagery during earnings events, and equity forecasting by intermediate fusion of market signals.

There is still a distinct difference between what machine learning research wants to do and what institutional investment needs to do. Most technical work is still mostly judged by prediction metrics. On the other hand, investment use cases also depend on calibrated uncertainty, resilience across market regimes, governance, and documentation that can stand up to review and audit. Current studies on multimodal learning effectively describe the subject from a technological standpoint. However, they don't fully address the practical limitations of using it for fiduciary investments. In contrast, model-risk and governance frameworks offer a framework for validation and supervision, but they do not directly address the added complexity introduced by multimodal systems.

This study fills that gap in two ways. First, it looks at the research on multimodal approaches used in investment-related applications and sums up the real-world data for different combinations of modalities. Second, it suggests the Multimodal Investment Readiness (MIR) paradigm to help the transition from research to institutional application. In this context, a more detailed explanation of three key concepts is needed: modality trust scoring, event-centric alignment contracts, and uncertainty-to-action mapping. These concepts aim to address two key inquiries: the influence of observable characteristics across various modalities on fusion design, and the impact of multimodal uncertainty on portfolio choices, particularly within the context of governance and regulatory frameworks.

2. Methodology

This review relies on a systematic synthesis of the literature. We chose sources from finance, computer science, and information systems journals that are indexed in Scopus, as well as conference papers that have been referenced a lot from ACL, NeurIPS, ICML, WWW, and CIKM. The search spanned from 2017 to early 2025, including combinations of the phrases “multimodal,” “fusion,” “representation learning,” and “finance,” along with domain-specific keywords such as “volatility,” “earnings calls,” “alternative data,” and “investment.” Baltrušaitis et al. (2019), Liang et al, Qin and Yang (2019), Yang et al. (2020), and Katona et al. (2024) have all made significant contributions to this field, along with Loughran and McDonald (2011) and the 2020 study, which also received considerable attention.

This review is divided into three main sections. It begins by describing different financial modalities and how they work. The subsequent discussion links the representation and fusion techniques to the specific requirements of financial applications. Ultimately, the paper's primary conceptual contribution, the MIR framework, is supported by an integration of empirical literature. Assertions are, wherever feasible, grounded in prior research. Furthermore, more cautious interpretations are meticulously elaborated upon.

3. The Intangible-Information Lens: Why Multimodality Matters for Finance

It is important to start the conversation with financial economics before looking at specific modalities and architectures. Investment results are influenced by information advantages that are intrinsically multimodal in character. A fundamental analyst who reads a 10-K file, listens to an earnings call, and looks at satellite images of shop traffic is doing implicit multimodal fusion. The efficient market theory (Fama, 1970) posits that prices should swiftly integrate all available information; nevertheless, empirical data repeatedly demonstrates that various information modalities are assimilated at differing velocities and efficiency.

Loughran and McDonald (2011) show that 73.8% of terms that general-purpose sentiment dictionaries say are negative are not negative in financial settings. This shows that even within a single modality (text), domain-specific representation is quite important. Katona et al.(2024) demonstrate that satellite-observable parking lot traffic reveals information regarding quarterly retailer performance that is not reflected in stock prices before public earnings announcements, indicating that markets process alternative data modalities even more slowly than text. Qin and Yang (2019) demonstrate that auditory signals in earnings calls convey information on future volatility that exceeds what is reflected in the written transcript alone. These results together indicate that multimodal representation learning, by concurrently encoding signals from many information modalities, may identify patterns that markets price inefficiently due to the isolated processing of modalities by human analysts and conventional models.

4. Modality Taxonomy for Financial Applications

A useful taxonomy must take into account both the kind of data and the operational factors that affect tradability and validation needs. We delineate five primary groups (Table 1). This taxonomy significantly expands upon previous classifications (Baltrusaitis et al., 2019) by incorporating operational dimensions—namely delay, revision behavior, and licensing constraints—that are crucial for deployment yet lacking in conventional multimodal surveys.

Table 1: Taxonomy of Financial Modalities with Operational Characteristics

Modality	Representative Signals	Typical Latency	Key Operational Pitfall
Time Series	Prices, volume, order flow, options surface	Milliseconds to daily	Non-stationarity, lookahead leakage
Text	10-K/10-Q filings, news, transcripts, social media	Minutes to quarterly	Narrative drift, 73.8% domain mismatch (Loughran & McDonald, 2011)
Audio	Earnings call prosody: pitch, jitter, shimmer, pauses	Post-call (hours)	Speaker variation, alignment error
Visual/Geospatial	Satellite imagery, foot traffic, industrial monitoring	Daily to weekly	Cloud obstruction, high cost, domain shift
Graphs/Relational	Supply chains, ownership, co-mention networks	Variable (constructed)	Entity resolution, temporal edge validity

The best multimodal use cases are those that focus on events. This is because different modes naturally sync up around profits, macro announcements, and geopolitical shocks, which makes it possible to align time and link causes to effects. To avoid lookahead bias when working with continuous signals at different frequencies, it's essential to model latency (de Prado, 2018).

5. Representation and Fusion Paradigms

5.1. Fusion Strategies and Empirical Evidence

There are three main types of fusion techniques that the literature talks about. Before encoding, early fusion combines raw characteristics. Lee and Yoo (2020) show that early fusion improves accuracy by 12% over unimodal baselines when utilizing cross-market equity data. Even so, it can overlook certain inputs. The most common way to achieve this is by intermediate fusion, which encodes modalities independently before integrating them using cross-attention, gating, or a mixture of these. Yang et al. (2020) indicate a 17-49% reduction in mean squared error (MSE) by hierarchical transformer intermediate fusion for predicting earnings call volatility. Late fusion ensembles provide predictions for each modality, which makes them more reliable but means they overlook interactions at the feature level (Baltrusaitis et al., 2019).

5.2. Contrastive Alignment

Contrastive aims synchronize modalities within common embedding spaces by attracting congruent cross-modal couples and rejecting incongruent ones. The CLIP model, developed by Radford and his team in 2021, created this framework for vision-language tasks. In finance, contrastive pretraining is appealing because to the inherent noise in supervised return labels, but cross-modal co-occurrence offers substantial self-supervision. Contrastive predictive coding, as described by van den Oord and his team in 2018, helps maintain temporal consistency, which is important for understanding forward-looking financial trends.

5.3. Transformer Architectures

Transformers (Vaswani et al., 2017) offer adaptable sequence modeling for text and time series (Wen et al., 2023). Cross-attention lets one kind of data ask a question of another kind of data. Yang et al. (2020) integrated token-level BERT encoding with sentence-level cross-modal attention to attain cutting-edge volatility forecasting. Sawhney et al. employed a multi-task architecture to improve forecasts of volatility and price fluctuations in a 2020 research presented at ACM Multimedia. The result was achieved by using specific text features related to the field and audio-based alignment.

6. Empirical Evidence from Finance-Specific Multimodal Models

This part puts together the empirical record, which is sorted by modality combination. Table 2 brings together the most important quantitative data.

Table 2: Selected Empirical Results from Multimodal Financial Models

Study	Venue	Modalities	Key Quantitative Finding
Qin & Yang (2019)	ACL	Text + Audio	MSE reduction: 6.5% (3-day) to 19.1% (30-day) vs. text-only
Yang et al. (2020)	WWW	Text + Audio	17-49% MSE improvement; multi-task learning enhances long-horizon
Li et al. (2020)	CIKM	Text + Audio	6x larger corpus (S&P 1500); sentence-level alignment
Sawhney et al. (2020)	ACM Multimedia	Text + Audio	Multi-task text-audio attention improves volatility and price movement prediction
Lee & Yoo (2020)	J. Supercomputing	Cross-market series	12-18% accuracy gain from intermediate fusion
Katona et al. (2024)	JFQA	Satellite + Prices	4.64% 3-day abnormal return around earnings announcements
Zhang et al. (2018)	Knowledge-Based Sys.	Text + Prices	Heterogeneous fusion outperforms single-source models
Zhu (2019)	Rev. Financial Studies	Satellite + Fundamentals	Satellite coverage increases stock price informativeness

6.1. Text-Audio Fusion: The Earnings Call Setting

Earnings calls are the most multimodal financial environment since they have text and audio that naturally fit together. In 2019, Qin and Yang introduced the multimodal deep regression model (MDRM) for analyzing S&P 500 companies. They show that audio signals (pitch, intensity, speech rate) may predict volatility better than text alone. Their MSE gains get better the longer the prediction horizon, going from 6.5% at 3 days to 19.1% at 30 days. This suggests that audio signals pick up on slow-moving uncertainty that text alone doesn't. Yang et al. (2020) made significant improvements using HTML design, achieving 17-49% reductions in mean squared error (MSE). This was done by combining hierarchical transformer encoding with multi-task learning to predict both volatility and price changes. The MAEC dataset, as described by Li et al. (2020), provides sentence-level alignment for companies in the S&P 1500. This dataset is six times larger than previous ones, allowing for more precise analysis. 6.2. Katona and colleagues (2024), in the Journal of Financial and Quantitative Analysis, present the strongest evidence from capital markets about alternative visual data. Their analysis of 4.8 million satellite data points from 67,120 U.S. retail locations shows that parking lot traffic signals can predict quarterly results, leading to an additional 4.64% increase in revenue during three-day announcement periods. This information isn't reflected in prices until it's publicly available, which is important because it could change established knowledge.

6.2. Satellite Imagery and Alternative Visual Data

Katona and colleagues (2024), in the Journal of Financial and Quantitative Analysis, present the most persuasive capital market evidence concerning alternative visual data. Their analysis of 4.8 million satellite data points from 67,120 U.S. retail locations reveals that parking lot traffic signals can forecast quarterly results, yielding an additional 4.64% in revenue during three-day announcement windows. This information is not reflected in prices until its public release, a crucial factor given its potential to alter long-standing knowledge. Furthermore, Zhu's (2019) research, published in The Review of Financial Studies, supports the idea that starting satellite coverage improves how well stock prices reflect information. This suggests that different perspectives function as true information primitives rather than mere approximations.

6.3. Domain-Specific Language Representations

How effectively the text is represented is a big part of how well multimodal systems work. Loughran and McDonald (2011) presented a research in The Journal of Finance that revealed that numerous terms that are designated as "negative" in general-purpose dictionaries don't mean the same thing in finance. They also discovered that 73.8% of the terms that the Harvard IV-4 lexicon said were bad were really neutral or even good in financial situations. This observation directly impacts multimodal systems, given that domain-agnostic text encoders frequently misinterpret financial sentiment, thereby disseminating inaccuracies through fusion layers. To mitigate this issue, FinBERT (Araci, 2019) modifies BERT for financial text, and BloombergGPT (Wu et al., 2023) illustrates the advantages of pretraining on a 363-billion-token financial corpus specifically designed for the domain.

7. The Multimodal Investment Readiness (MIR) Framework

The preceding review reveals a systematic disconnect between multimodal ML research and institutional investment requirements. Research papers optimize prediction accuracy on held-out data, while institutional deployment demands calibrated uncertainty, missing-modality robustness, regulatory auditability, and explicit decision policies. To bridge this gap, we propose the Multimodal Investment Readiness (MIR) framework, which introduces three novel concepts and maps observable modality features through intangible information channels to concrete deployment requirements.

7.1. Modality Trust Scoring

Existing multimodal architectures typically treat all modalities as equally reliable at inference time or apply static dropout during training. In production investment systems, modality reliability is non-stationary and regime-dependent: news feeds may become unreliable during flash crashes, satellite imagery suffers seasonal cloud obstruction, and earnings call audio quality varies across companies. We formalize modality trust scoring as a learned, time-varying reliability weight computed from observable modality health indicators (latency deviation, completeness ratio, distributional anomaly score) that feeds directly into the fusion gating mechanism.

Formally, for modality m at time t, the trust score is defined as tau_m(t) = g(latency_m(t), completeness_m(t), anomaly_m(t); theta), where g is a learned function parameterized by theta, trained jointly with the fusion model to minimize prediction loss while penalizing overreliance on low-trust modalities. This differs from standard modality dropout in three ways: (i) trust scores are conditioned on real-time observable quality metrics rather than random masking, (ii) trust scores vary continuously rather than binary on/off, and (iii) trust score trajectories themselves become a monitorable production metric for model risk governance. When a modality's trust score drops below a learned threshold, the system gracefully degrades rather than producing unreliable predictions.

7.2. Event-Centric Alignment Contracts

Timestamp misalignment is the most prevalent leakage channel in multimodal finance (de Prado, 2018). Yet existing datasets and models rarely formalize alignment rules as first-class architectural components. We introduce event-centric alignment contracts: formal specifications defining, for each modality pair and event type, (i) the temporal window of permissible data inclusion relative to event time, (ii) the data availability delay that must be respected, and (iii) the handling protocol for missing or revised data.

Alignment contracts serve a dual function. Technically, they constrain the data pipeline to prevent leakage by construction rather than relying on post-hoc verification. Regulatorily, they constitute auditable governance artifacts satisfying SR 11-7 requirements for data lineage documentation (Board of Governors of the Federal Reserve System, 2011). For instance, an alignment contract for analyzing earnings calls might say that text transcripts can only be made available after the filing time (usually T+1 business day for 8-K), audio features can only be made available after the call end timestamp, and price labels can only be calculated from T+2 open to keep information from being added overnight. This contract is now part of the model documentation package that is looked over during independent validation.

7.3. Uncertainty-to-Action Mapping

The third MIR component deals with the difference between estimating uncertainty (which has been well-studied technically in the past) and taking investment action, which means turning uncertainty into position sizes, hedge ratios, and decisions about when to defer.We suggest that multimodal investing systems ought to have a distinct uncertainty-to-action layer that translates calibrated joint uncertainty, including both per-modality and fusion-level uncertainty, into portfolio choices.

This layer implements three action channels. Position sizing is inversely related to prediction uncertainty, which means that when model confidence is low, exposure is lower. As tail uncertainty (CVaR of the uncertainty distribution) rises, hedging intensity rises as well. This makes sure that protective positions grow in line with negative risk estimate. When joint uncertainty goes over a specific level that is set based on the institution's risk tolerance, deferral kicks in. This means that the decision will go to a human for review instead of being made automatically.

Table 3: The MIR Framework: Mapping Modality Features to Deployment Requirements via Information Channels

Observable Feature	Information Channel	MIR Component	Deployment Requirement
Modality latency, completeness, anomaly rate	Signal reliability	Modality Trust Scoring	Graceful degradation; production monitoring
Event timestamps, filing times, data availability delays	Temporal integrity	Alignment Contracts	Leakage prevention; SR 11-7 data lineage
Prediction variance, ensemble disagreement, conformal intervals	Decision confidence	Uncertainty-to-Action Mapping	Position sizing, hedging, deferral policies
Per-modality contribution to prediction	Interpretabilit y	Trust Score + Attention Weights	Explainability for governance review
Cross-regime performance stability	Robustness	All three components jointly	Stress testing; regime-conditional validation

The MIR framework is intentionally conceptual rather than prescriptive about specific architectures. You may use the trust score system with cross-attention gating, mixture-of-experts routing, or dedicated reliability networks. You may enforce alignment contracts by using data pipeline limitations or formal verification. Any calibrated uncertainty approach can be used by the uncertainty-to-action layer. The framework's usefulness comes from making these criteria clear and giving a clear path from research prototype to a production system that can be audited.

8. Evaluation Methodology: Why Results Often Do Not Replicate

8.1. Backtest Overfitting

Multimodal systems broaden the hypothesis space via modalities, alignment parameters, and fusion hyperparameters, hence increasing the risk of overfitting. Bailey et al. (2017) in the Journal of Computational Finance present a formal methodology for calculating the probability of back test overfitting (PBO) and demonstrate that the usual hold-out method is inaccurate for investment simulations. For a strong assessment, you need walk-forward validation, purged cross-validation with embargo (de Prado, 2018), and adjustments for repeated testing.

8.2. Decision-Relevant Metrics

Table 4 contrasts standard metrics used in ML papers with decision-relevant alternatives required for institutional deployment.

Table 4: . Standard ML Evaluation Metrics and Alternatives Focused on Decision-Making

Dimension	Common ML Evaluation Practice	More Decision-Oriented Measure
Accuracy	MSE, MAE, and F1-score are commonly reported	Measures such as Sharpe ratio, Sortino ratio, and information ratio better reflect decision usefulness
Risk assessment	Risk is often not explicitly evaluated	Max drawdown, CVaR, and tail ratio provide a clearer view of downside exposure
Cost sensitivity	Trading or operational costs are often omitted	Turnover and transaction-cost analysis help assess real-world feasibility
Robustness	Performance is often tested on only one train-test split	Walk-forward testing, purged cross-validation, and regime-based analysis offer stronger robustness checks
Uncertainty estimation	Prediction uncertainty is frequently left unreported	ECE, Brier score, and abstention curves help evaluate confidence and calibration
Temporal validity	Random splitting is commonly used for validation	Point-in-time data and embargo-based validation better preserve temporal realism

9. Governance and Regulatory Considerations

SR 11-7 (Board of Governors of the Federal Reserve System, 2011) mandates that material models undergo documentation, independent validation, and continuous monitoring. The NIST AI RMF 1.0 (NIST, 2023) furnishes a structure for risk identification, while the EU AI Act (European Parliament, 2024) specifies sector-specific regulations for AI systems with financial ramifications.. Governance of multimodal systems necessitates the management of modality data lineage, the execution of alignment integrity audits (facilitated by MIR alignment contracts), and the implementation of robustness testing. The MIR framework's trust scores and alignment contracts function as governance mechanisms: trust score trajectories furnish auditable records of modality reliance, and alignment contracts serve as verifiable specifications of temporal integrity during independent validation.

10. Framework Illustration: Applying MIR to Earnings Call Analysis

To demonstrate the practical application of the MIR framework, we trace its three components through a concrete multimodal investment pipeline: predicting post-earnings-call volatility from text, audio, and price data.

10.1. Alignment Contract Specification

For an S&P 500 company's quarterly earnings call, which takes place at market close on day T, the alignment contract outlines the following: (a) **Price data**: Only closing prices up to and including day T-1 are utilized. The T-day close is omitted to prevent any potential information leak stemming from intraday market reactions to the call. (b) The text transcript becomes accessible following the 8-K filing. For the majority of companies, this occurs on T+1, with the SEC EDGAR filing timestamp marking the official start. (c) Audio features: accessible after the call is over and only if the conference provider sends the call end timestamp. (d) Labels: volatility calculated from T+2 to T+N, leaving out T and T+1 to show the difference between the prediction and the immediate announcement effect. This contract is stored as a version-controlled artifact, checked during independent model validation according to SR 11-7, and enforced in the data pipeline through code.

10.2. Trust Scoring in Practice

During the call analysis, each modality receives a trust score. For text, the trust function monitors transcript completeness (partial vs. full), filing timeliness (delays exceeding the historical distribution trigger trust reduction), and distributional deviation of embedding norms from the training set. For audio, trust monitors recording quality (signal-to-noise ratio), speaker diarization confidence, and forced alignment accuracy. For price data, trust monitors liquidity (abnormally low volume degrades trust), halt indicators, and microstructure anomalies. When a company reports during a market disruption event (for example, a flash crash), the price modality trust score drops automatically, and the fusion module upweights text and audio contributions without requiring manual intervention.

10.3. Uncertainty-to-Action Mapping

The system produces a volatility forecast with calibrated prediction intervals. The uncertainty-to-action layer then maps this to portfolio decisions. If the 90% prediction interval for 7-day realized volatility is [15%, 22%], the system sizes positions inversely proportional to interval width. The system will not run if the interval is wider than the 95th percentile of previous interval widths, for example. Instead, it will wait for a human to examine it. Automatic hedging happens when the lower bound of the interval is higher than the portfolio's risk budget. We use conformal prediction on validation data to set these thresholds so that we can be guaranteed of finite-sample coverage guarantees (Angelopoulos & Bates, 2021).We then reset them every three months as part of the model monitoring cycle.

10.4. Note on Source Tiers

In the interest of scholarly transparency, we note that the sources cited in this review span three tiers: (a) peer-reviewed journals indexed in Scopus (e.g., The Journal of Finance, JFQA, Review of Financial Studies, IEEE TPAMI, Knowledge-Based Systems, Journal of Supercomputing, Journal of Computational Finance, International Journal of Machine Learning and Cybernetics); (b) top-tier peer-reviewed conference proceedings indexed in Scopus and widely recognized in computer science (ACL, WWW, CIKM, ACM Multimedia, NeurIPS, ICML); and (c) highly cited preprints not yet published in journals (arXiv papers by Liang et al., 2022; Wu et al., 2023; Angelopoulos & Bates, 2021; Araci, 2019).In computer science, conference proceedings are just as prestigious as journal articles in domains like finance. However, we want to make this clear for readers who come from financial backgrounds, where journal articles are the main quality signal. All of the arXiv-only sources are referenced for basic technical ideas, and they all have more than 500 citations, which shows that the community accepts them.

11. Limitations

This paper is not an actual research; it is a conceptual review and framework suggestion. The MIR framework has not been tested by being used on real trading systems. The modality trust score system is based on theory, but it has to be tested in different market conditions and asset classes. In latency-sensitive situations, alignment contracts may add extra work and costs that make them less useful. The uncertainty-to-action mapping necessitates institution-specific threshold calibration.Furthermore, the literature reviewed is weighted toward equities and may not generalize to fixed income, commodities, or private markets.

12. Conclusion

Multimodal representation learning provides increasingly mature tools for investment systems that mustintegrate heterogeneous information, particularly around financially material events. The empirical evidence is strongestfor text-audio fusion in earnings call settings, with models achieving 17-49% MSE improvements (Yanget al., 2020), and for satellite imagery as an alternative data modality, with demonstrated alpha of 4.64% around earnings events (Katona et al., 2024).

However, research-stage performance alone is insufficient for institutional deployment. This paper contributes the MIR framework, which formalizes three concepts bridging the gap between multimodal ML research and investment practice: modality trust scoring enables dynamic, regime-conditional reliability weighting; event-centric alignment contracts provide leakage prevention and governance documentation simultaneously; and uncertainty-to-action mapping translates model confidence into concrete portfolio decisions. These parts work together to create a clear path from research prototype to high-stakes deployment that can be audited.

Future research should empirically validate the MIR framework across various asset classes and market regimes, establish standardized benchmarks that integrate the evaluation dimensions identified in this review, and investigate how extensive multimodal foundation models trained on financial corpora may transform the modality taxonomy and fusion architecture landscape.

Reference

1. AIMA. (2020). Casting the net: How hedge funds are using alternative data. Alternative Investment Management Association.
2. Angelopoulos, A. N., & Bates, S. (2021). A gentle introduction to conformal prediction and distribution-free uncertainty quantification. arXiv:2107.07511.
3. Araci, D. (2019). FinBERT: Financial sentiment analysis with pre-trained language models. arXiv:1908.10063.
4. Bailey, D. H., Borwein, J. M., Lopez de Prado, M., & Zhu, Q. J. (2017). The probability of backtest overfitting. Journal of Computational Finance, 20(4), 39-69.
5. Baltrusaitis, T., Ahuja, C., & Morency, L.-P. (2019). Multimodal machine learning: A survey and taxonomy. IEEE Transactions on Pattern Analysis and Machine Intelligence, 41(2), 423-443.
6. Board of Governors of the Federal Reserve System. (2011). SR 11-7: Supervisory guidance on model risk management.
7. de Prado, M. L. (2018). Advances in financial machine learning. John Wiley & Sons.
8. European Parliament. (2024). Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (AI Act). Official Journal of the European Union.
9. Fama, E. F. (1970). Efficient capital markets: A review of theory and empirical work. The Journal of Finance, 25(2), 383-417.
10. Gal, Y., & Ghahramani, Z. (2016). Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In Proceedings of ICML (pp. 1050-1059).
11. Guo, C., Pleiss, G., Sun, Y., & Weinberger, K. Q. (2017). On calibration of modern neural networks. In Proceedings of ICML (pp. 1321-1330).
12. Katona, Z., Painter, M., Patatoukas, P. N., & Zeng, J. (2024). On the capital market consequences of alternative data: Evidence from outer space. Journal of Financial and Quantitative Analysis, 59(4), 1-29.
13. Lakshminarayanan, B., Pritzel, A., & Blundell, C. (2017). Simple and scalable predictive uncertainty estimation using deep ensembles. In Advances in NeurIPS, 30.
14. Lee, S. I., & Yoo, S. J. (2020). Multimodal deep learning for finance: Integrating and forecasting international stock markets. The Journal of Supercomputing, 76, 8294-8312.
15. Li, J., Yang, L., Smyth, B., & Dong, R. (2020). MAEC: A multimodal aligned earnings conference call dataset for financial risk prediction. In Proceedings of CIKM (pp. 3063-3070). ACM.

16. Liang, P. P., Zadeh, A., & Morency, L.-P. (2022). Foundations and trends in multimodal machine learning: Principles, challenges, and open questions. arXiv:2209.03430.
17. Loughran, T., & McDonald, B. (2011). When is a liability not a liability? Textual analysis, dictionaries, and 10-Ks. The Journal of Finance, 66(1), 35-65.
18. National Institute of Standards and Technology. (2023). Artificial intelligence risk management framework (AI RMF 1.0). U.S. Department of Commerce.
19. Ngiam, J., Khosla, A., Kim, M., Nam, J., Lee, H., & Ng, A. Y. (2011). Multimodal deep learning. In Proceedings of ICML (pp. 689-696).
20. Qin, Y., & Yang, Y. (2019). What you say and how you say it matters: Predicting stock volatility using verbal and vocal cues. In Proceedings of ACL (pp. 390-401).
21. Radford, A., Kim, J. W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., ... & Sutskever, I. (2021). Learning transferable visual models from natural language supervision. In Proceedings of ICML (pp. 8748-8763).
22. Sawhney, R., Mathur, P., Mangal, A., Khanna, P., Shah, R. R., & Zimmermann, R. (2020). Multimodal multi-task financial risk forecasting. In Proceedings of ACM Multimedia (MM) (pp. 456-465). ACM.
23. van den Oord, A., Li, Y., & Vinyals, O. (2018). Representation learning with contrastive predictive coding. arXiv:1807.03748.
24. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. In Advances in NeurIPS, 30.
25. Wen, Q., Zhou, T., Zhang, C., Chen, W., Ma, Z., Yan, J., & Sun, L. (2023). Transformers in time series: A survey. International Journal of Machine Learning and Cybernetics, 14, 3681-3703.
26. Wu, S., Irsoy, O., Lu, S., Daber, V., Dredze, M., Gehrmann, S., ... & Mann, G. (2023). BloombergGPT: A large language model for finance. arXiv:2303.17564.
27. Yang, L., Ng, T. L. J., Smyth, B., & Dong, R. (2020). HTML: Hierarchical transformer-based multi-task learning for volatility prediction. In Proceedings of WWW (pp. 441-451). ACM.
28. Zhang, X., Zhang, Y., Wang, S., Yao, Y., Fang, B., & Yu, P. S. (2018). Improving stock market prediction via heterogeneous information fusion. Knowledge-Based Systems, 143, 236-247.
29. Zhu, C. (2019). Big data as a governance mechanism. Review of Financial Studies, 32(5), 2021-2061.