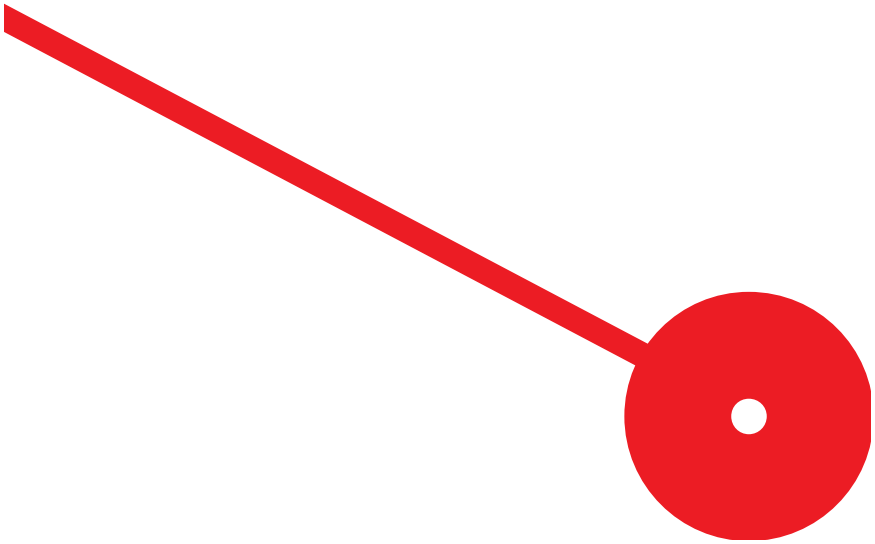
INSTITUTO
SUPERIOR
DE CONTABILIDADE
E ADMINISTRAÇÃO
DO PORTO
POLITÉCNICO
DO PORTO

MASTER'S DEGREE
BUSINESS INTELLIGENCE AND ANALYTICS

Multimodal Fusion for Time Series Forecasting: Learning from Temporal and Visual Data

Alexandre Rafael Seabra Bento

10/2025



Alexandre Rafael Seabra Bento. Multimodal Fusion for Time Series Forecasting: Learning from Temporal and Visual Data 10/2025

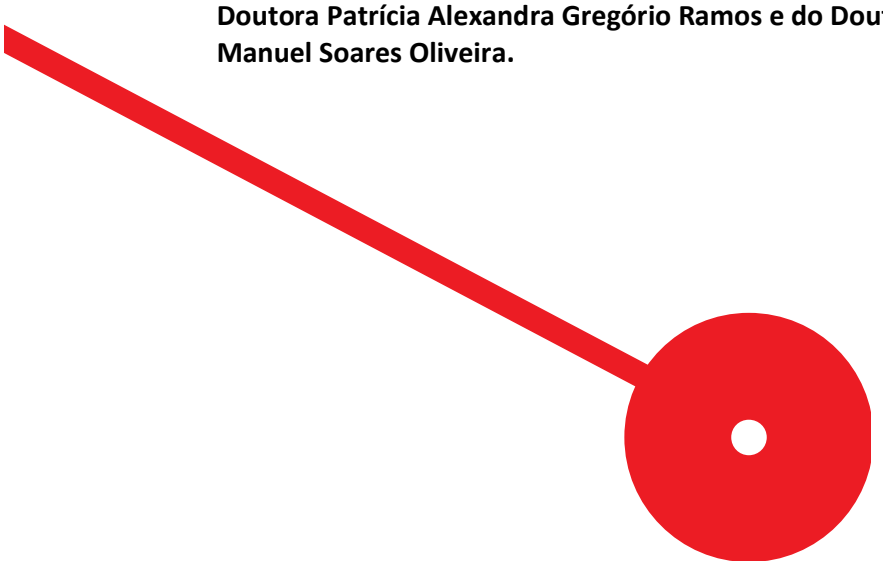
INSTITUTO
SUPERIOR
DE CONTABILIDADE
E ADMINISTRAÇÃO
DO PORTO
POLITÉCNICO
DO PORTO

MASTER'S DEGREE
BUSINESS INTELLIGENCE AND ANALYTICS

Multimodal Fusion for Time Series Forecasting: Learning from Temporal and Visual Data

Alexandre Rafael Seabra Bento

Dissertação de Mestrado apresentada ao Instituto Superior de Contabilidade e Administração do Porto para a obtenção do grau de Mestre em Business Intelligence and Analytics, sob orientação da Doutora Patrícia Alexandra Gregório Ramos e do Doutor José Manuel Soares Oliveira.



Alexandre Rafael Seabra Bento. Multimodal Fusion for Time Series Forecasting: Learning from Temporal and Visual Data
10/2025

Dedication

To my family, whose unwavering support, patience, and encouragement have been the foundation of all my achievements.

To my parents, for their lifelong example of perseverance and integrity.

To my friends, for their understanding and companionship throughout this journey.

And to all those who, through their guidance and belief in me, have made this accomplishment possible.

Acknowledgements

This thesis would not have been possible without the support of many individuals and institutions, to whom I express my sincere gratitude.

First and foremost, I thank Dr. Patrícia Alexandra Gregório Ramos and Dr. José Manuel Soares Oliveira for their scientific supervision, academic rigor, continuous availability, and constructive guidance, which substantially enhanced the quality of this work. Their methodological rigor, attention to detail, and trust were decisive throughout the project.

I also thank Instituto Superior de Contabilidade e Administração do Porto and the Master's in Business Intelligence and Analytics for a demanding and stimulating academic environment, and all faculty members for the knowledge and competencies provided during the program. The support of the administrative services was equally important for the smooth progress of this journey.

My appreciation extends to classmates and project peers for idea sharing, productive discussion, and a strong spirit of collaboration, which helped refine methodological choices and validate results.

I am grateful to professionals and partners who, directly or indirectly, facilitated access to tools, references, and best practices relevant to Multimodal Fusion for Time Series Forecasting, informing the literature review, experimental procedures, and discussion of findings.

Finally, I thank my family and friends for their unwavering support, understanding, and encouragement throughout this process. Their patience and confidence were vital.

To all who contributed, explicitly or behind the scenes, thank you.

Resumo

Os modelos de previsão de séries temporais que recorrem apenas a dados numéricos tendem a ignorar fatores exógenos e padrões estruturais que podem ser mais facilmente captados por representações visuais. Esta dissertação propõe um enquadramento multimodal que integra sequências temporais numéricas e representações visuais (*plots*/imagens) para melhorar a exatidão, a robustez e a interpretabilidade da previsão. Metodologicamente, emprega-se um FT-Transformer para a componente temporal e uma rede convolucional TIMM para a componente visual, combinadas por um esquema de fusão híbrida a meio da rede. A *pipeline* inclui normalização e padronização de *plots*, geração consistente de janelas temporais, otimização bayesiana com Optuna e protocolos de avaliação reprodutíveis.

A avaliação é realizada no subconjunto de séries com frequência horária do conjunto de dados M4, em múltiplos horizontes (1–48 passos), reportando o NRMSE agregado e estratificado (1–12, 13–24, 25–36, 37–48). Os resultados mostram que o modelo multimodal supera consistentemente as variantes unimodais (apenas numéricas e apenas visuais), com melhorias até 7,0% em NRMSE face ao melhor *baseline*, enquanto estudos de ablação evidenciam o contributo específico do ramo visual e do mecanismo de fusão.

Contribui-se, assim, com: (i) um *framework* multimodal eficiente e reprodutível; (ii) um protocolo experimental transparente para fusão numérico-visual; e (iii) diretrizes práticas sobre normalização de *plots*, janelas e *tuning*. Discutem-se limitações, como a sensibilidade ao estilo do gráfico e à sincronização temporal, e traçam-se direções futuras que incluem a integração de texto contextual e a previsão sensível a intervenções, visando sistemas de previsão mais adaptativos e aplicáveis ao mundo real.

Palavras chave: Séries Temporais; Aprendizagem Multimodal; Fusão Numérico-Visual; Previsão (NRMSE)

Abstract

Time series forecasting models that rely solely on numerical data often overlook exogenous factors and structural patterns that can be more effectively captured through visual representations. This thesis proposes a multimodal framework that integrates numerical time series sequences and visual representations (plots/images) to enhance forecasting accuracy, robustness, and interpretability. Methodologically, the approach employs an FT-Transformer for temporal processing and a TIMM-based convolutional network for visual feature extraction, combined through a hybrid mid-level fusion strategy. The training pipeline includes normalization and standardization of plots, consistent generation of temporal windows, Bayesian hyperparameter optimization with Optuna, and reproducible evaluation protocols.

The framework is evaluated on the hourly subset of the M4 dataset, across multiple forecasting horizons (1–48 steps), reporting both aggregated and stratified Normalized Root Mean Squared Error (NRMSE) metrics (1–12, 13–24, 25–36, 37–48). Results demonstrate that the multimodal model consistently outperforms unimodal variants (numerical-only and visual-only), achieving up to 7.0% NRMSE reduction compared to the best baseline, whereas ablation studies highlight the specific contribution of the visual branch and the fusion mechanism.

This research contributes: (i) an efficient and reproducible multimodal forecasting framework; (ii) a transparent experimental protocol for numerical–visual fusion; and (iii) practical guidelines on plot normalization, window generation, and model tuning. Limitations, such as sensitivity to plot style and temporal synchronization, are discussed, along with future directions including the integration of contextual text and intervention-aware forecasting for more adaptive, real-world prediction systems.

Keywords: Time Series; Multimodal Learning; Numerical–Visual Fusion; Forecasting (NRMSE)

Table of Contents

Table of Figures

Table of Tables

List of abbreviations

AR - Autoregressive

ARFIMA - Autoregressive Fractionally Integrated Moving Average

ARIMA - Autoregressive Integrated Moving Average

ARMA - Autoregressive Moving Average

BPNN - Backpropagation Neural Network

CiK - Context-in-Knowledge (multimodal integration strategy based on context-aware retrieval)

CNN - Convolutional Neural Network

DeepAR - Deep Autoregressive Recurrent model

DLinear / NLinear - Decomposed Linear / Nonlinear Linear models for long-term series forecasting

EHR - Electronic Health Records

EMD - Empirical Mode Decomposition

EMA - Exponential Moving Average

FEDformer - Frequency Enhanced Decomposed Transformer

FT-Transformer - Feature Tokenizer Transformer (temporal encoder for tabular or time series inputs)

FX - Foreign Exchange

GARCH - Generalized Autoregressive Conditional Heteroskedasticity

GPT4MTS - Generative Pre-trained Transformer for Multivariate Time Series (LLM adapted to temporal data)

Grad-CAM - Gradient-weighted Class Activation Mapping

GPU - Graphics Processing Unit

iTransformer - Inverted Transformer (variable-wise Transformer architecture for time series forecasting)

Informer - Efficient Transformer model for long sequence time series forecasting

InfoNCE - Info Noise-Contrastive Estimation (contrastive learning objective)

I/O - Input/Output

LEMMA-RCA - Large Multi-modal Multi-domain Dataset for Root Cause Analysis

LLM - Large Language Model

LSTM - Long Short-Term Memory

LTSF - Long-Term Series Forecasting

M4 - M4 Forecasting Competition dataset

MA - Moving Average

MAE - Mean Absolute Error

MASE - Mean Absolute Scaled Error

MEDHMP - Medical Hierarchical Multimodal Pretraining

MLP - Multilayer Perceptron

MMF - Multimodal Fusion (integration of heterogeneous modalities such as text, image, and time series)

MM-TSF - Multimodal Time Series Forecasting

MM-TSFlib - Multimodal Time Series Forecasting Library (benchmarking framework for multimodal forecasting)

MoLE - Mixture-of-Linear-Experts

MSE - Mean Squared Error

NRMSE - Normalized Root Mean Squared Error

Optuna - Hyperparameter optimization framework

PatchTST - Patch-based Time Series Transformer

RNN - Recurrent Neural Network

RMSE - Root Mean Squared Error

sMAPE - Symmetric Mean Absolute Percentage Error

SVR - Support Vector Regression

SVM - Support Vector Machine

TaTS - Text-attributed Time Series (dataset type combining text descriptions with numeric series)

TGTSF - Text-Guided Time Series Forecasting

TiDE – Time series Dense Encoder

TimeCAP – Time series Context-Aware Prompting (method integrating prompts into temporal modeling)

Time-MMD – Time series Multimodal Distillation (knowledge distillation framework for multimodal forecasting)

TIMM - PyTorch Image Models (library for pretrained CNN visual backbones)

TimesNet - Temporal 2D-Variation modeling network for time series analysis

xTP-LLM - Cross-Temporal Prompted Large Language Model (LLM-based framework for temporal reasoning)

CHAPTER I – INTRODUCTION

1 Introduction

Time series forecasting plays a pivotal role in contemporary decision-making across domains such as finance, energy, healthcare, and transportation. It provides the analytical foundation for anticipating future dynamics from historical data, thereby supporting planning, optimization, and risk management. Over the past decade, the field has been profoundly shaped by advances in deep learning architectures - most notably recurrent neural networks, temporal convolutional models, and, more recently, Transformers - each capable of capturing complex temporal dependencies that traditional statistical methods often fail to model. Nevertheless, most of these models continue to operate under a unimodal paradigm, learning exclusively from numerical sequences. They assume that the past numerical trajectory contains all the relevant information required to predict the future. In practice, this assumption rarely holds true. Many of the drivers behind temporal variations such as policy interventions, market sentiment, contextual events, or operational changes - are exogenous to the observed signal and cannot be inferred solely from numeric history.

This gap between observable numeric trends and the hidden contextual mechanisms governing them has led to the emergence of multimodal time series forecasting (MTSF) as a coherent and increasingly mature research direction. MTSF treats heterogeneous evidence - numerical sequences, textual information, plots, images, and even spatiotemporal signals - as complementary perspectives on the same underlying process. Rather than relying solely on the numeric evolution of a variable, multimodal approaches attempt to integrate external or latent cues that shape its dynamics. This integration can occur across different stages of the modeling pipeline - at the feature level (early fusion), representation level (intermediate fusion), or decision level (late fusion) - and aims to enhance both predictive accuracy and interpretability.

The motivation for this shift is grounded in how human analysts reason about time-dependent data. In practical business and scientific contexts, experts rarely interpret long numeric columns in isolation. They typically rely on visual representations, annotations, and contextual information to discern structure, regimes, and anomalies. A financial analyst, for instance, might identify trend reversals more easily from the shape of a price trajectory plotted over time than from raw values alone. Similarly, operational planners frequently contextualize numeric data with external narratives such as policy

announcements or event logs. These human patterns of multimodal reasoning suggest that numerical data, when paired with visual or textual cues, may allow learning systems to generalize more effectively and exhibit better robustness to noise, missing values, and structural breaks.

Despite this intuitive premise, multimodal forecasting remains an evolving and heterogeneous field. Most progress to date has occurred along textual-temporal fusion, in which large language models (LLMs) are used to interpret or augment numeric sequences. Recent studies have shown that integrating texts - such as news, social media, or event reports - can improve the ability of models to anticipate regime changes and explain prediction shifts. Some approaches employ LLM agents that engage in iterative reasoning cycles of prediction, critique, and refinement, while others generate textual embeddings aligned with numeric timelines, creating what has been called chronological textual resonance. These models embody a movement towards context-aware forecasting systems that treat text not merely as auxiliary metadata, but as an integral part of the causal and semantic fabric underlying the series.

In contrast, visual modality has received comparatively little systematic attention, even though evidence increasingly indicates that visual representations can encode structural information that numeric-only models fail to capture. The visual form of a plot inherently conveys macro-level regularities - such as seasonality, periodicity, and overall shape - that complement the fine-grained temporal detail encoded in numerical values. Moreover, visual features may act as inductive priors, guiding the model towards global consistency and enhancing interpretability by providing a common visual language shared by human analysts and algorithms alike. Early studies on multimodal fusion have suggested that “plots unlock understanding” in multimodal time series models, yet there is still no consensus on when and how visual features provide measurable benefits, nor on the best strategies to normalize, synchronize, or fuse these modalities in a consistent way.

This thesis addresses that gap by proposing and evaluating a multimodal fusion framework that jointly learns from numerical and visual representations of the same temporal process. Specifically, it introduces a forecasting model that combines a temporal encoder based on the FT-Transformer - responsible for capturing local and sequential dependencies - with a visual encoder based on TIMM/MetaFormer (CaFormer-B36), trained on standardized plots of the same historical window. The embeddings produced

by both encoders are projected into a shared latent space, where an intermediate fusion mechanism aggregates complementary information before feeding a regression head tasked with predicting a multi-step horizon of 48 future values. This design aims to merge local temporal structure with global visual context, leveraging the strengths of both modalities under a single, unified architecture.

The methodological contribution of this work extends beyond architecture design. It proposes a reproducible and efficient experimental protocol encompassing plot normalization, consistent windowing (960 input steps → 48 forecast steps), Bayesian hyperparameter optimization using Optuna, mixed-precision training, and rigorous cross-run reproducibility through fixed random seeds. This level of transparency is critical for ensuring fair comparison across unimodal and multimodal variants. The model is evaluated on the hourly subset of the M4 dataset, a well-established benchmark for time series forecasting, using the Normalized Root Mean Squared Error (NRMSE) as the primary metric. Results are reported both in aggregate across the full forecast horizon (1–48) and stratified by contiguous blocks (1–12, 13–24, 25–36, 37–48), allowing a fine-grained analysis of horizon-dependent behavior.

Empirical results demonstrate that the proposed multimodal model consistently outperforms its unimodal counterparts - both numeric-only and visual-only - achieving up to a 7,0% reduction in NRMSE relative to the best baseline. Furthermore, multimodal fusion leads to lower error dispersion across test series and horizons, suggesting enhanced stability and generalization. Ablation experiments confirm that these gains cannot be attributed solely to architectural complexity: when either the visual or temporal branch is removed, performance degrades substantially, and disabling the fusion layer nullifies the advantage entirely. Visual modality contributes particularly to the stability of long-horizon forecasts, acting as a global regularizer that mitigates local noise and smoothing inconsistencies, while the temporal encoder ensures short-term precision. Together, these findings validate the central hypothesis that numeric–visual complementarity is both measurable and practically exploitable in forecasting tasks.

This thesis also aims to situate this contribution within a broader landscape of multimodal research. Recent advances in intervention-aware forecasting have shown that explicitly modeling external shocks or decisions can reduce structural error ceilings, transforming forecasting from a purely statistical exercise into a dynamic systems problem. Meanwhile, LLM-based reasoning systems have reintroduced contextual and

causal understanding into time series analysis, enabling models to reflect on their outputs or incorporate unstructured textual information. Foundational multimodal models, such as those designed for paired time–text data, have further demonstrated the potential for zero-shot transfer across domains. By focusing on the visual branch of this growing multimodal ecosystem, this work provides a complementary and reproducible perspective: it examines how standardized visualizations - devoid of external data - can themselves constitute an informative and robust modality for forecasting.

Beyond accuracy, this research emphasizes interpretability and practicality. Visual saliency maps and attention mechanisms are used to identify which temporal segments or visual features most influence the final predictions, providing transparency suitable for business applications where explainability is paramount. The controlled normalization of plots also allows the study of style sensitivity, ensuring that observed improvements are due to representational content rather than aesthetic artifacts. These aspects contribute to a model design that is not only effective but also auditable, a crucial step towards the responsible integration of multimodal systems in decision support contexts.

The scope of this study is deliberately focused on numeric and visual modalities derived from the same source data. This constraint ensures that improvements can be attributed to multimodal fusion rather than to the injection of external information. Nonetheless, several limitations are acknowledged. The model's sensitivity to visual style and temporal synchronization remains an open challenge, and the current evaluation is confined to a single benchmark (M4-hourly). Future research may extend the framework to include textual and event-based context, adaptive fusion strategies per horizon, and causal interpretability components. Ethical and methodological transparency remain guiding principles, particularly given the growing influence of generative and multimodal models in domains with significant societal or financial impact.

In summary, this thesis contributes to the emerging field of multimodal time series forecasting by demonstrating, through a rigorously controlled and reproducible study, that visual information is not merely aesthetic but structurally informative for temporal prediction. By fusing numerical and visual embeddings within a unified framework, it offers empirical evidence that multimodal integration enhances both predictive performance and model stability, while laying the groundwork for broader multimodal

architectures capable of incorporating language, interventions, and higher-order contextual reasoning in the future.

2 State of the art

Multimodal time series forecasting (MTSF) has matured into a coherent research field that regards heterogeneous evidence - numerical sequences, textual narratives, visual plots and images, spatiotemporal graphs, and contextual signals - as complementary manifestations of the same underlying generative processes. The unifying premise is that many real-world drivers of temporal change lie outside the numeric signal itself and therefore remain invisible to unimodal forecasters. Jiang et al. (2025) consolidate this view through a comprehensive tutorial and survey that organize multimodal time series analysis along three interacting design axes - fusion, alignment, and transference - emphasizing that practical systems must handle heterogeneity in sampling rates, temporal misalignment, missing or noisy modalities, and domain shift, while retaining interpretability and computational tractability. Within this framework, fusion may occur at the feature, intermediate, or decision level, often realized through attention mechanisms; alignment concerns synchronization and event-lag modeling across modalities; and transference covers pre-training and adaptation strategies that enable robust performance across datasets and domains. The present chapter situates this thesis within that expanding landscape, focusing especially on recent advances in cross-modal reasoning and the integration of visual, textual, and numerical cues for enhanced forecasting accuracy, robustness, and interpretability.

2.1 Early statistical foundations (2004-2010)

The first half of the 2000s was dominated by models that extended the Box-Jenkins methodology into new contexts. Pai and Lin (2005) proposed one of the most cited early hybrid frameworks that combined the autoregressive integrated moving-average (ARIMA) model with support vector machines (SVMs). Recognizing ARIMA's strength in modeling linear autocorrelations and SVM's ability to capture nonlinear residual structures, they introduced an approach that first filtered the data through ARIMA and then learned the residual component with SVM regression. Their experiments on stock-price data revealed that the hybrid scheme outperformed both pure statistical and pure neural approaches, demonstrating that linear and nonlinear dynamics could coexist within financial series.

At roughly the same time, Bhardwaj and Swanson (2006) investigated autoregressive fractionally integrated moving-average (ARFIMA) processes. Their

empirical analysis provided one of the earliest systematic tests of long-memory behavior in macroeconomic and financial returns. By comparing ARFIMA against AR, MA, ARMA, and Generalized Autoregressive Conditional Heteroskedasticity (GARCH) benchmarks, they found that fractional differencing could improve multi-step forecasts for equity indexes when long-range dependence was evident, although short-horizon prediction gains were modest. This work established the theoretical basis for later interest in models capable of handling both persistence and volatility clustering.

In the latter part of the decade, research interest expanded to nonparametric and semi-functional formulations. Aneiros-Pérez and Vieu (2008) introduced a semi-functional partial-linear model that treated each historical trajectory as a continuous function rather than a discrete vector. By embedding functional predictors into a partially linear regression structure, they enabled the simultaneous capture of smooth functional effects and linear covariate influences. This approach marked an early attempt to move beyond fixed-parameter models toward data-adaptive representations of temporal structure.

2.2 Emergence of hybrid and nonlinear models (2011-2015)

By the early 2010s, the limitations of purely linear statistical models had become apparent. Scholars began merging traditional time series methods with neural computation and evolutionary optimization. Wang et al. (2012) proposed a hybrid exponential-smoothing-ARIMA-back-propagation neural network (BPNN) optimized with a genetic algorithm. Their framework exploited exponential smoothing and ARIMA to learn linear and seasonal components, while the BPNN captured remaining nonlinearities. Using data from the Shenzhen and Dow Jones indexes, they demonstrated consistent accuracy gains relative to individual component models and even to equal-weight or random-walk ensembles. This study exemplified a transitional phase in which statistical decomposition and neural learning were explicitly combined.

Parallel progress occurred in the broader econometrics' literature. Phan et al. (2015) reassessed the predictability of stock returns using out-of-sample forecasting tests. They examined how oil-price dynamics affected sectoral equity returns and found that predictive power was strongly sector-dependent. Their conclusion - that traditional predictive regressions often fail out of sample - highlighted the need for more flexible, data-driven approaches capable of adapting to regime changes.

During the same period, ARIMA remained a reference baseline. Mondal et al. (2014) applied it to fifty-six Indian equities across industries, exploring how the choice of historical window length influenced forecast accuracy. Despite its simplicity, their work reaffirmed ARIMA's robustness and interpretability, establishing benchmarks for later nonlinear models.

Another strand of research sought to incorporate exogenous, text-derived variables. Rather than relying solely on prices and indicators, contemporaneous work showed that coupling nonlinear autoregressive architectures with sentiment time series extracted from financial news or social media can materially enhance prediction of equity indexes-anticipating the multimodal and sentiment-augmented frameworks that would flourish a decade later (Anbaee Farimani et al., 2022; Liapis et al., 2023).

2.3 The deep learning turn (2016–2020)

2.3.1 Decomposition and kernel methods

Entering the late 2010s, researchers increasingly targeted the non-stationarity and multi-scale nature of financial data. Nava et al. (2018) proposed a multistep-ahead methodology that fused empirical mode decomposition (EMD) with support vector regression (SVR). EMD separated high and low-frequency intrinsic-mode functions, each of which was modeled with SVR. The ensemble of components yielded superior forecasts of the S&P 500 index at horizons from 30 seconds to 25 minutes. The authors emphasized that high-frequency modes contributed to short-term accuracy, whereas low-frequency modes captured longer-term trends, thereby directly addressing the twin challenges of non-stationarity and multi-scaling in financial time series.

2.3.2 Recurrent neural networks and LSTM extensions

The success of deep learning in other domains soon reached forecasting. Sagheer and Kotb (2019) introduced a deep long short-term memory (DLSTM) architecture optimized via a genetic algorithm for petroleum-production forecasting. Their results, validated on real oilfield data, showed that the DLSTM surpassed both shallow recurrent networks and classical statistical models. By stacking multiple LSTM layers, the network captured dependencies over extended horizons, signaling a paradigm shift toward hierarchical deep representations of temporal processes. In the same period, fully connected residual stacks demonstrated that non-recurrent backbones could also excel:

Oreshkin et al. (2019) proposed N-BEATS, an interpretable MLP architecture with basis expansions that rivaled classical methods while offering decomposition-style insights.

2.3.3 Convolutional-recurrent hybrids

Hybridization continued with convolutional-recurrent designs. Vidal and Kristjanpoller (2020) developed a CNN-LSTM model for gold-volatility forecasting. Their approach combined convolutional layers-effective in extracting local spatial patterns from images of volatility series with LSTM units that modeled long-term sequential dependencies. Compared with GARCH and standalone LSTM baselines, their hybrid achieved a 37% reduction in mean squared errors. The study not only demonstrated deep learning's superiority in capturing complex dynamics but also introduced image-based feature representations into time series prediction.

2.3.4 Reappraising classical models in the big-data era

Despite these deep learning advances, classical models retained importance as benchmarks and interpretable baselines. Studies continued to revisit ARIMA's utility in practice and to use it as a strong point of comparison, reaffirming that, under data limitations or tight interpretability constraints, well-specified linear models can remain competitive (Mondal et al., 2014; Pai and Lin, 2005).

2.4 Synthesis of the 2004–2020 period

Over this sixteen years period, three evolutionary threads have become apparent. First, hybridization emerged as the central strategy for bridging statistical rigor and nonlinear adaptability. From the ARIMA-SVM architecture of Pai and Lin (2005) to the EMD-SVR and CNN-LSTM frameworks of the late 2010s (Nava et al., 2018; Vidal and Kristjanpoller, 2020), researchers repeatedly combined complementary modeling philosophies rather than replacing one with another. Second, there was a gradual broadening of data modalities; sentiment integration in financial markets foreshadowed the data-fusion trend that would later dominate multimodal forecasting (Anbaee Farimani et al., 2022; Liapis et al., 2023). Finally, the rise of deep learning marked a structural change in how temporal dependencies were represented: rather than specifying lag orders or difference parameters manually, models learned hierarchical or residual representations directly from data (Sagheer and Kotb, 2019; Oreshkin et al., 2019). By 2020, forecast research had shifted decisively from deterministic parameterization toward

automatic feature learning, setting the stage for Transformer architecture and foundation-model paradigms.

2.5 Transformer-based and hybrid deep models (2021–2023)

The early 2020s witnessed an acceleration of Transformer architecture - originally designed for natural-language processing - into time series forecasting. Their self-attention mechanisms promised to overcome the vanishing-gradient limitations of recurrent networks while capturing long-range temporal dependencies more efficiently. Beyond sequence-to-sequence attention, frequency-aware and tokenization-aware designs expanded the toolkit, and linear-centric reappraisals introduced strong baselines.

2.5.1 Decomposition-enhanced and frequency-aware transformers

Decomposition-enhanced pipelines and frequency-domain modeling improved long-horizon stability. Zhou et al. (2021)'s Informer introduced ProbSparse attention for efficient long-context modeling, while Zhou et al. (2022)'s FEDformer decomposed signals and modeled components in the frequency domain, strengthening low-frequency representation. Wu et al. (2023) proposed TimesNet, reframing 1D sequences as 2D temporal-variation maps to capture multi-periodicity with CNN-like inductive biases. Nie et al. (2023) advanced tokenization with PatchTST, which uses subseries-level patches and channel independence to encode local semantics without sacrificing global context.

2.5.2 Sentiment-enhanced financial prediction

The surge in language models prompted renewed attention to sentiment as an exogenous predictor (Anbaee Farimani et al., 2022) combined FinBERT-derived sentiment with technical indicators in deep recurrent pipelines, reducing errors in FX and crypto. Complementarily, broad benchmarks showed that, when aligned carefully, sentiment time series can bolster intermediate-horizon accuracy (Liapis et al., 2023).

2.5.3 Deep sentiment fusion and feature diversity

In systematic comparisons across thirty algorithms and sixty-seven feature suites, Liapis et al. (2023) demonstrated that sentiment-price fusion benefits hinge on the temporal alignment and quality of text-derived signals, highlighting the importance of leak-aware preprocessing and horizon-wise evaluation.

2.5.4 CNN-transformer fusion for financial sequences

For intraday S&P 500 forecasts, Zeng et al. (2023) coupled CNNs for local temporal motifs with Transformers for global context, surpassing ARIMA, EMA, and DeepAR baselines and illustrating the complementarity between locality and long-range attention.

2.5.5 The linear resurgence and efficient MLP backbones

Challenging Transformer hegemony, Zeng et al. (2023) showed that deceptively simple decomposed linear models (DLinear/NLinear) can outperform sophisticated attention models on standard LTSF suites. Das et al. (2023) proposed TiDE, a dense MLP encoder that achieved competitive accuracy with notable speed benefits. Building on this, Ni et al. (2024) introduced Mixture-of-Linear-Experts (MoLE), which ensembles lightweight linear experts with learned routing to recover flexibility while retaining efficiency. These results underscore that inductive bias and evaluation protocol often matter more than raw architectural complexity.

2.6 Multimodal and pretraining frameworks (2023-2024)

As data sources diversified, researchers pursued architectures capable of learning from multiple modalities - numerical, textual, and visual.

2.6.1 Hierarchical multimodal pretraining in healthcare

Beyond finance, Wang et al. (2023) presented the Medical Hierarchical Multimodal Pretraining (MEDHMP), a hierarchical pretraining framework for multimodal Electronic Health Records (EHRs). By integrating temporal vital-sign sequences with categorical and textual clinical information through multi-level self-supervision, MEDHMP illustrated how pretraining unifies heterogeneous time-dependent data and informs cross-domain forecasting design.

2.6.2 Semi-functional and nonlinear extensions

In parallel, functional-data ideas regained relevance as interpretable complements to deep networks. Studies building on semi-functional partial-linear modeling (Aneiros-Pérez and Vieu, 2008) showed that functional embeddings can be integrated into modern architectures to preserve transparency while benefiting from nonlinear representation learning.

2.7 Foundation models and the inversion of transformers (2024–2025)

The mid-2020s brought the notion of foundation models - large, pretrained networks transferable across forecasting tasks - into the time series domain, alongside architectural reframings of attention and tokenization.

2.7.1 Inverted transformers for forecasting

Liu et al. (2024) proposed iTransformer, reframing the Transformer by inverting its attention dimension: variables become tokens, attention models cross-variate structure, and per-variate temporal mapping shifts to feed-forward blocks. This achieved state-of-the-art performance across diverse datasets while scaling gracefully to long lookbacks.

2.7.2 Large-scale, multi-domain foundation forecasting

Foundation pretraining has progressed along two complementary lines. Language-model-style tokenization of scaled values enables probabilistic forecasting in a decoder-only setting (Ansari et al., 2024; Das et al., 2024), delivering strong zero-shot performance across heterogeneous datasets. In a different vein, tabular foundation models can transfer to time series without temporal pretraining when paired with lightweight timestamp features: Hoo et al. (2025) adapted TabPFN-v2 to forecasting (TabPFN-TS) and achieved top leaderboard results in a zero-shot regime. For classification, Wang et al. (2025) reformulated multivariate time series as training-free table understanding (TableTime), leveraging LLM reasoning on serialized tables to obtain robust zero-shot accuracy.

2.7.3 Cross-domain and spatially grounded LLMs

Beyond purely temporal settings, LLMs are being grounded in spatial context and diagnostic signals. Feng et al. (2025) used city-scale instruction-tuning and self-weighted fine-tuning to enhance urban spatial cognition, while Zheng et al. (2025) released the Large Multi-modal Multi-domain Dataset for Root Cause Analysis (LEMMA-RCA), a large multi-modal multi-domain dataset (metrics + logs) for root-cause analysis that stress-tests models under abrupt shifts. These resources broaden the evaluation of multimodal reasoning and indicate how forecasting, detection, and explanation can be unified within foundation-model ecosystems.

2.8 Cross-modal design space: fusion, alignment, and robustness

The design space synthesized by Jiang et al. (2025) and refined in subsequent works (Daswani et al., 2024; Liu et al., 2025; Kim et al., 2024; Williams et al., 2025; Xu et al., 2025; Jiang et al., 2025; Liu et al., 2025) converges on several key dimensions of model design. Fusion strategies have evolved from simple feature concatenation, which is easy to implement but susceptible to variance dominance, toward intermediate fusion via cross-attention or lightweight adapter layers, which balance modality interaction and stability. Late decision-level ensembling remains relevant for robustness when some modalities are unavailable at inference. Alignment continues to be a decisive factor, as practical systems must reconcile mismatched sampling rates, publication lags, and differing horizon structures. Common solutions include lag-search algorithms, timestamp-filtered retrieval, and attention-based temporal alignment mechanisms. Robustness is addressed through drop-modality training, mixture-of-experts architectures with learned gating functions, and cross-modal distillation, which allows unimodal student models to inherit multimodal teacher performance. Representation learning leverages self-supervised pretraining: temporal encoders benefit from masked-token or contrastive objectives, vision encoders from masked-image modeling and temporally coherent augmentations, and multimodal encoders from cross-modal contrastive learning objectives such as InfoNCE. Finally, evaluation protocols have expanded to include not only conventional accuracy metrics such as NRMSE, sMAPE, and MASE but also interpretability diagnostics, horizon-wise performance reports, and controlled experiments that introduce synthetic misalignment or missing modalities. These practices, now widely adopted following Jiang et al. (2025), form the empirical and methodological foundation of contemporary multimodal forecasting.

2.9 Modalities and representations

2.9.1 Text-guided forecasting and event alignment

Textual information, though unstructured and noisy, provides rich context that can greatly enhance predictive power when temporally aligned with numerical signals. Xu et al. (2025) introduced Text-Guided Time Series Forecasting (TGTSF), using cross-attention to fuse news and descriptive metadata with numeric histories. Subsequent works, such as the one from Wang et al. (2024) (From News to Forecast), refine this approach by employing LLM-based agents that iteratively filter out irrelevant narratives

and reflect upon forecasting errors to update reasoning logic. These reflective agents yield more coherent integrations of event semantics and temporal causality. Jia et al. (2024) extend the paradigm by integrating Google Trends, news, and sales data, showing that explicitly monitoring overfitting-generalization trade-offs sustains fusion robustness. Jiang et al. (2025) categorize alignment strategies into four major techniques: resampling and lag-search to reconcile sampling mismatches, attention-based soft correspondences between text spans and time tokens, timestamp-filtered retrieval to prevent future leakage, and noise-gating or drop-modality training for resilience. These mechanisms directly inform the design of both text-series and plot-series fusion pipelines.

2.9.2 Text as time series and context augmentation

Several recent contributions reconceptualize text itself as a time-structured signal. Li et al. (2025) formulate Chronological Textual Resonance (CTR) within the Texts-as-Time-Series (TaTS) framework, revealing periodicities in textual data that mirror numerical seasonality. By embedding texts at each timestamp and quantifying their spectral alignment via a TT-Wasserstein metric, they demonstrate that paired texts can act as auxiliary variables enhancing forecasting and imputation. Lee et al. (2025) propose TimeCAP, where specialized LLM agents contextualize time series behavior by summarizing recent patterns before prediction. Their framework, which combines contextualization, augmentation, and prediction agents, achieves substantial improvements across healthcare, climate, and finance domains. Complementary to these, Liu et al. (2025) release the Time-MMD corpus, a multi-domain multimodal dataset pairing numeric series with timestamped textual context; it establishes standardized evaluation for text-series integration and advocates reproducible tooling. Collectively, these studies confirm Jiang et al. (2025)'s argument that progress in multimodal forecasting depends on high-quality aligned datasets, transparent benchmarks, and modular frameworks capable of exploiting temporal structure in language itself.

2.9.3 Visual modalities and plot-centric reasoning

Visual evidence is emerging as a crucial complement to numeric signals. Two principal families of visual modalities dominate the literature: raw scene or remote-sensing imagery that encodes physical state and plot-based representations that visualize temporal evolution as charts or figures. Daswani et al. (2024) demonstrate that simply rendering time series data as plots and feeding them into multimodal foundation models

allows vision transformers to “see” temporal structure directly. Their experiments reveal up to 150% performance gains and 90% cost reductions relative to textual encoding, validating that plots convey high-density temporal semantics accessible to vision encoders. This finding aligns with Jiang et al. (2025), who explain that patch-wise attention over charts captures global shape and long-range dependencies, while local convolutions stabilize motif extraction. Self-supervised pre-training through masked-image modeling on domain-specific plots further improves robustness to noise and reduces labeled-data requirements. To prevent spurious visual cues, the authors recommend standardizing axes, scales, and plotting styles and temporally anchoring each plot window to the same look-back horizon used by the temporal encoder. Saliency-mapping techniques such as Grad-CAM can then expose the model’s focus areas, producing human-readable rationales such as pre-holiday regime shifts or anomaly peaks. Complementary work by Guo et al. (2024) in transportation forecasting supports the same intuition: when multimodal large-language-model (LLM) frameworks translate heterogeneous sensor data into structured textual or visual prompts, interpretability improves without degrading accuracy. Collectively, these results suggest that plot-centric multimodal reasoning is not merely a visualization convenience but a legitimate representational channel.

2.9.3.1 Vision encoders as temporal abstractions

The broader trend of employing vision transformers for time series reasoning has motivated new encoder architectures where temporal data are reinterpreted as images or patches. This approach complements the FT-Transformer backbone introduced by Gorishniy et al. (2023), which tokenizes tabular or numerical inputs for attention-based learning. When combined with vision embeddings, these transformers capture both local feature-level interactions and global temporal dependencies. The synergy between visual abstraction and token attention thus becomes a key design principle in multimodal forecasting pipelines.

2.10 Architecture and modeling strategies

2.10.1 LLM-native multimodal forecasters

A strong research current treats time series as a “foreign language,” allowing the direct use of LLM backbones. Wang et al. (2024) introduce ChatTime, a unified multimodal time series foundation model that encodes numeric data as discrete tokens

analogous to words and supports bimodal input/output (text ↔ series). ChatTime achieves zero-shot forecasting, context-guided prediction, and time series question answering through shared tokenization and instruction-tuning. Parallel work by Jia et al. (2024) (GPT4MTS) employs prompt engineering and retrieval-augmented generation (RAG) to inject relevant historical context, substantially improving cross-domain generalization. Merrill et al. (2024) caution, however, that even advanced language models still struggle to reason about temporal dynamics in zero-shot settings: in controlled tests, humans outperform GPT-4 by up to 30 percentage points in etiological and question-answering tasks. To mitigate this gap, Jiang et al. (2025) recommend explicit temporal tokenization, calendar embeddings, and iterative reasoning loops (predict → critique → revise). The collective evidence points toward a synthesis of linguistic and temporal modeling, where large-language foundations provide general reasoning capability, while temporal encoders contribute precision and stability.

2.10.2 Joint prediction of time series and text

Beyond treating text as auxiliary input, several systems pursue joint multimodal generation. Kim et al. (2024) propose the Multi-Modal Forecaster, which predicts both numerical trajectories and textual summaries, compelling the model to internalize cross-modal dynamics rather than regard text as static metadata. The approach connects to emerging reasoning-capable forecasters such as Time-LLM and Chronos foundations, where consistency between numeric predictions and generated textual rationales serves as an implicit regularizer. Williams et al. (2025) analyze these models through the Context is Key (CiK) benchmark, arguing that evaluating whether the model actually "reads" the contextual language is essential. CiK's tasks require text comprehension for successful forecasting, transforming simple correlation matching into text-grounded reasoning about interventions and regime shifts.

2.10.3 Intervention-aware and system-level perspectives

Classical forecasting assumes stationarity, but real-world systems are subject to exogenous shocks. Xu et al. (2025) propose Intervention-Aware Forecasting (IATSF), grounded in control theory, to incorporate external textual interventions explicitly. They introduce FIATS, a lightweight model with Channel-Aware Adaptive Sensitivity Modeling (CASM) and Channel-Aware Parameter Sharing (CAPS), enabling channel-specific responses to interventions. Their theoretical analysis shows that ignoring

interventions introduces an irreducible error bound - formal proof of the self-stimulation limitation. By modeling interventions directly, IATSF reframes forecasting as dynamic-system inference under external control, aligning with causal reasoning frameworks and providing interpretable sensitivity maps. This system-level view complements event-based textual integration (Wang et al., 2024) and supports more robust evaluation under distributional shifts.

2.10.4 Reasoning and interpretability pipelines

Interpretability has become a primary design objective. Jiang et al. (2025) introduced TimeXL, an explainable multimodal prediction system coupling a prototype-based encoder with three LLM agents - prediction, reflection, and refinement. This closed-loop workflow (predict → reflect → refine) improves both accuracy and human-centric explanations. Kim et al. (2024)'s Hybrid-MMF further demonstrates that naive feature concatenation rarely yields gains; effective fusion requires gating and modulation. Williams et al. (2025) show that explicit prompts or retrieved context often outperform complex joint embeddings when textual cues are decisive. Liu et al. (2025) explore cognitive System 1 vs System 2 reasoning for zero-shot tasks, finding that deliberative reasoning disproportionately benefits multimodal settings. Together, these works crystallize a design pattern where fast pattern extractors (temporal and visual encoders) are coupled with slow reasoning modules (LLM controllers) to balance efficiency and interpretability.

2.10.5 Operational explainability: prototypes, sensitivity, and plots

Explainability that practitioners can audit requires bridging numeric, textual, and visual rationales. Jiang et al. (2025) operationalize this through prototype-based reasoning, while Xu et al. (2025)'s CASM formulates channel-specific sensitivity maps grounded in control-theoretic semantics. Daswani et al. (2024) emphasize that even visual rendering choices affect interpretability: differing plot styles can shift model attention, underscoring the need for standardized visualization pipelines. These techniques collectively establish an auditable interpretability chain from input modalities to forecast explanations, aligning with the transparency standards advocated in broader generative-AI governance research (Changalidis et al., 2025).

2.10.6 LLM-native routes versus structured fusion

Prompt-native systems such as GPT4MTS (Jia et al., 2024) treat text as adaptive soft prompts over frozen backbones, gaining efficiency but relying on careful retrieval and prompt engineering. Structured-fusion models (e.g., cross-attention or adapter-based hybrids) (Jiang et al., 2025; Kim et al., 2024; Jiang et al., 2025) offer greater stability and explicit feature exchange. The consensus emerging across studies (Jiang et al., 2025; Liu et al., 2025; Jia et al., 2024; Merrill et al., 2024; Kim et al., 2024; Williams et al., 2025; Xu et al., 2025) is that intermediate-level cross-attention fusion remains the most effective compromise between interpretability, robustness, and computational cost.

2.11 Areas of application

2.11.1 Transportation forecasting and decision-making with explainability

Transportation systems exemplify the challenges of multimodal forecasting, as they reflect both endogenous periodicities and exogenous shocks from events, weather, or infrastructure changes. Liao et al. (2022) introduce MIFGNN, a multimodal graph neural model that fuses temporal and contextual evidence - including weather, events, and spatial graphs - through attention-based message passing. Their experiments on taxi demand prediction reveal that cross-modal contributions vary across forecasting horizons, emphasizing the importance of adaptive gating mechanisms to prevent modality dominance. In parallel, Guo et al. (2024) present xTP-LLM, which converts heterogeneous spatial-temporal inputs into structured textual prompts consumable by LLMs. This method delivers not only accurate forecasts but also interpretable natural-language explanations, thus bridging predictive modeling and decision support. Collectively, these studies confirm the principle advocated by Jiang et al. (2025) that human-centered interpretability is not a secondary objective but a prerequisite for operational uptake in intelligent transportation systems.

2.11.2 Economics and finance: alignment as a first-order effect

Economic and financial time series are highly sensitive to external events and news, making them ideal domains for exploring alignment effects in multimodal forecasting. Mou et al. (2025) propose MM-iTransformer, a multimodal architecture that fuses textual embeddings derived from domain-specific encoders such as FinBERT with temporal sequences of prices and indicators. By inverting the conventional attention structure to

treat variates as tokens and time as features, their model captures cross-asset dependencies such as lead-lag relationships and dynamic cointegration. Ablation studies demonstrate that accurate temporal alignment between news and target windows is essential: misaligned news causes severe performance degradation, and domain-specific encoders consistently outperform general-purpose LLMs. Jiang et al. (2025) generalize these observations, noting that event conditioning offers the greatest improvements during volatile regimes and that interpretable diagnostics - such as attention heatmaps or top-k explanatory news snippets - are indispensable for analyst trust in financial forecasting.

2.12 Data sets, benchmarks, and evaluation

2.12.1 Multimodal datasets for forecasting

The lack of standardized multimodal datasets has long impeded progress. Liu et al. (2025) address this gap with Time-MMD, pairing numeric series with timestamped textual context across nine domains. The dataset incorporates leak-free filtering, summary generation, and domain labeling, and ships with MM-TSFlib for baseline evaluation. Time-MMD's design philosophy - alignment, disentanglement, and transparency - has influenced later corpora such as the Temporal-Synced IATSF benchmark (Xu et al., 2025). Jiang et al. (2025) emphasize that datasets combining numeric streams with essential textual information enable evaluation under realistic exogenous conditions, whereas unimodal baselines miss the influence of policy or news events entirely.

2.12.2 Benchmarks with essential textual information

Williams et al. (2025)'s Context is Key (CiK) benchmark complements Time-MMD by ensuring that textual context is decisive for task success. CiK spans 71 tasks over seven domains and introduces the Region-of-Interest CRPS (RCRPS) metric to penalize forecasts that ignore textual constraints. By construction, no model can solve CiK tasks without understanding the text, creating a rigorous testbed for context-aware reasoning. The benchmark further mitigates data contamination through live-data sourcing and synthetic variations, establishing a new standard for evaluating whether models use rather than memorize context.

2.12.3 Metrics, leakage control, and "context necessity"

Evaluation design strongly influences reported gains. CiK's RCRPS (Williams et al., 2025) quantifies context-sensitivity by focusing on windows where text imposes

constraints. Jiang et al. (2025) recommend leak-aware curation pipelines including timestamp filtering and lag-search diagnostics. Xu et al. (2025) further synchronize textual interventions to patch boundaries, eliminating subtle future leakage. These studies collectively advocate explicit leak checks during dataset construction, metrics that penalize ignoring textual cues, and ablations that deliberately break alignment to measure sensitivity, ensuring that multimodal improvements are genuinely causal rather than artefactual.

2.12.4 Cost/latency trade-offs and token budgets

Computation efficiency increasingly shapes model design. Williams et al. (2025) show that Direct Prompt with strong LLMs yields top accuracy but high latency, while smaller specialized models narrow the gap at lower cost. Daswani et al. (2024) demonstrate that visual plots compress long numeric histories into compact vision tokens, achieving better functional-form recognition with drastically reduced token budgets. This finding supports the notion that the representational compression of visual encodings can outperform purely textual serialization, particularly for long-horizon temporal structures. A practical approach therefore emerges in which lightweight numeric forecasters serve as the fast default, and more computationally expensive LLM or vision conditioning modules are selectively activated only when the forecasting task requires contextual reasoning or cross-modal structure to capture extended dependencies.

2.12.5 Data breadth versus decisiveness: complementary benchmarks

Time-MMD (Liu et al., 2025) provides broad coverage across multiple domains and modalities, while CiK (Williams et al., 2025) enforces context necessity by constructing cases where textual or exogenous information is indispensable. When used together, these two benchmarks establish a dual experimental regime: Time-MMD enables general pretraining and large-scale ablation studies, whereas CiK stress-tests the model's ability to genuinely exploit context. This duality mirrors the distinction between pretraining-centric foundation models such as ChatTime (Wang et al., 2024) and reasoning-centric frameworks such as ReC4TS (Liu et al., 2025). Integrating both types of benchmarks supports a holistic evaluation pipeline for multimodal forecasting systems, balancing data diversity with diagnostic rigor.

2.13 Reasoning strategies for zero-shot multimodal forecasting

Reasoning augmentation represents one of the most promising directions for multimodal forecasting. Liu et al. (2025) introduced ReC4TS, the first benchmark designed to evaluate reasoning strategies such as self-consistency, deliberation, and reflection across unimodal and multimodal settings. Their experiments demonstrate that test-time self-consistency improves accuracy and that multimodal models benefit disproportionately from reasoning enhancements because textual and visual signals activate more structured inferential chains. Jiang et al. (2025) operationalize this idea in TimeXL through an explicit predict-critique-refine loop, while Wang et al. (2024)'s From News to Forecast integrates agentic reflection to dynamically refine event selection and reasoning logic. Merrill et al. (2024) complement these findings by revealing the persistent reasoning gap between humans and LLMs in zero-shot scenarios. Collectively, these works suggest that multi-stage reasoning mechanisms are not ancillary but fundamental to achieving robust generalization and interpretability in multimodal forecasting.

2.14 Positioning of the present work

The literature consistently demonstrates that fusing textual and numerical modalities improves forecasting accuracy when alignment and filtration are carefully managed, yet the incorporation of visual modalities remains relatively underexplored. Plot-centric learning, as shown by Daswani et al. (2024), provides compact and information-rich representations of temporal dynamics that vision transformers can exploit. Jiang et al. (2025) highlight that vision encoders excel at capturing global temporal structures and long-range dependencies, while self-supervised pretraining on standardized plot corpora enhances noise resilience. The domain-specific insights from MM-iTransformer (Mou et al., 2025) further emphasize that cross-modal alignment and adaptive gating mechanisms are critical, particularly when integrating visual encoders with temporal backbones. The present thesis directly addresses this research gap by designing a hybrid architecture in which a visual encoder (MetaFormer/TIMM) processes standardized plots synchronized to the same look-back horizon used by an FT-Transformer temporal encoder. The normalized embeddings from both are projected into a shared latent space and fused via an intermediate attention block that learns cross-modal interactions while regularizing for robustness. Empirical analysis evaluates horizon-wise

performance, sensitivity to misalignment, and resilience under missing-modality scenarios, complemented by visual attributes that reveal which structural cues drive the model's predictions.

2.15 Practical blueprint derived from the literature

Synthesizing the methodological insights from Jiang et al. (2025), Daswani et al. (2024), Liu et al. (2025), and Xu et al. (2025) yields a coherent experimental blueprint for multimodal forecasting. Data preparation should standardize the generation and scaling of plots while ensuring temporal anchoring to the same look-back window as the numeric input. Textual documents should be windowed into temporally consistent segments with explicit lag-search to capture publication and reaction delays. Temporal encoders benefit from masked-token or contrastive pretraining over the dataset, while visual encoders benefit from masked-image pretraining with augmentations that preserve temporal semantics. Fusion is most effective at the intermediate level, using learned projections and normalization, and should incorporate drop-modality training so the model degrades gracefully when inputs are missing or corrupted. Evaluation should combine NRMSE, sMAPE, and MASE metrics with horizon-wise analyses, ablations isolating visual fusion effects, synthetic misalignment tests, and robustness evaluations under modality dropout. Interpretability must be integral to the design, with attention maps over time tokens and saliency maps on plots that highlight the features influencing predictions. Collectively, these practices translate the general principles established across the literature into a concrete and reproducible methodological framework for visual and temporal fusion in applied business forecasting.

2.16 Synthesis: toward text-grounded, reasoned, and intervention-aware forecasting

Taken together, the reviewed corpus indicates that the next generation of forecasting systems will need to curate temporally aligned textual and visual context under strict leakage control (Liu et al., 2025; Williams et al., 2025), jointly optimize objectives that predict both numeric trajectories and explanatory language (Kim et al., 2024), incorporate explicit reasoning strategies such as self-consistency and reflection (Liu et al., 2025; Jiang et al., 2025), and evaluate under intervention-aware regimes that capture causal dependencies (Xu et al., 2025). This vision complements the multimodal economic modeling of MM-iTransformer (Mou et al., 2025) and the unified foundation

modeling of ChatTime (Wang et al., 2024), outlining a roadmap from treating text and visuals as auxiliary features toward fully text-grounded and causally reasoned forecasting paradigms.

2.17 Broader backdrop: applications and governance

Beyond algorithmic progress, responsible deployment of multimodal and generative-AI systems demands transparency and governance. Changalidis et al. (2025) provide a systematic review of generative-AI applications in genomics, emphasizing interpretability, validation, and ethical oversight in multimodal systems that combine structured and unstructured data. Although their domain differs, the same principles apply to forecasting in business and economics: interpretable visualizations, explicit rationales, and sensitivity analyses are prerequisites for the integration of multimodal forecasting into operational decision pipelines. These governance-oriented frameworks reinforce that explainability is not only a research goal but also an ethical and regulatory requirement for trustworthy AI-driven forecasting systems.

3 Methodology

This chapter presents the methodological framework developed for the multimodal time series forecasting model proposed in this study (see Figure 1). The methodology was designed to combine numerical and visual data representations within an integrated learning architecture, allowing the model to extract complementary information from both modalities. The goal of this approach is to enhance forecasting accuracy, robustness, and interpretability by leveraging the quantitative precision of temporal data together with the structural insights inherent in visual patterns. The following sections describe, in detail, the datasets used, the architecture of the proposed model, the training and optimization process, and the evaluation procedures adopted to assess its performance.

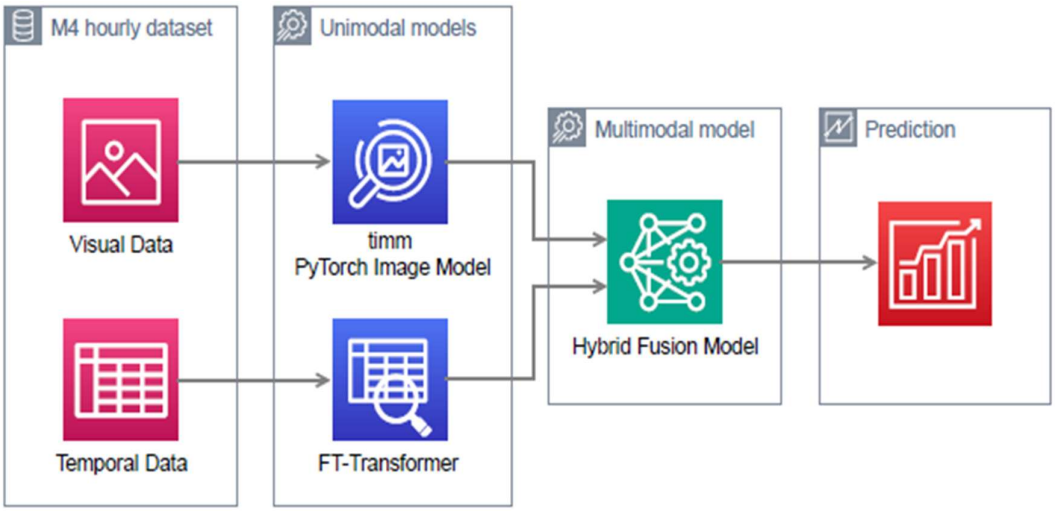


Figure 1. Multimodal architecture employed for time series forecasting.

3.1 Problem framing and benchmark data

The forecasting problem is defined as a supervised regression task, in which future values of a time series are predicted based on two sources of information: the numerical sequence of past observations and its corresponding visual representation. This multimodal formulation assumes that numerical and visual modalities offer distinct yet complementary perspectives on the underlying temporal dynamics. Numerical data captures detailed quantitative fluctuations, whereas visual plots encode higher-level

contextual information such as trends, periodicities, and anomalies that are not easily discernible in purely numeric form.

To guarantee comparability and robustness, the framework was evaluated on the M4 benchmark, specifically using the hourly subset, which contains a large and heterogeneous collection of real-world time series across multiple application domains. A total of 245 hourly series were selected, with 210 allocated for training and 35 reserved for testing. Each series consisted of 960 hourly observations serving as the input window, and the subsequent 48 observations were used as the forecasting horizon. This configuration allows for a balanced evaluation of short-, medium-, and long-term prediction capabilities, following standard practices in time series forecasting.

3.1.1 Data preparation and preprocessing

Prior to modeling, all series were preprocessed to ensure temporal alignment and numerical consistency. For the visual modality, each 960-step input window was rendered as a time series plot that retained the true temporal scale and range of values. The plots were standardized using consistent axis limits, aspect ratios, and color palettes to prevent the model from learning from stylistic differences rather than temporal patterns. Each image was resized and normalized according to the visual encoder's input specifications, while the numerical data were normalized on a per-series basis to stabilize training and facilitate convergence (see Figure 2). This preprocessing ensured that both modalities provided harmonized and noise-free input representations for subsequent encoding.

3.2 System overview and end-to-end pipeline

The overall system was implemented as an end-to-end pipeline that automates all stages of the modeling process, including data preparation, training, validation, hyperparameter optimization, and performance evaluation. This automation ensures reproducibility, scalability, and consistency across experiments.

The framework operates under three different configurations: a temporal-only model that uses numeric data exclusively, a visual-only model that relies solely on image-based inputs, and a multimodal model that fuses both modalities. The multimodal configuration forms the core of the proposed approach. In this design, temporal sequences are processed by a Transformer-based encoder, visual plots are analyzed by a convolutional encoder following the MetaFormer paradigm, and the resulting

representations are fused through a hybrid mechanism before producing the final forecasts. This integrated structure allows the model to learn both fine-grained temporal dependencies and high-level contextual features simultaneously.

3.3 Temporal modality

The numerical modality is modeled using a Feature-Tokenizer Transformer (FT-Transformer), chosen for its ability to capture long-range temporal dependencies and complex seasonal behavior in time series data. The FT-Transformer first embeds the numeric inputs into a high-dimensional space, followed by three stacked Transformer blocks. Each block contains a multi-head self-attention mechanism with eight attention heads, enabling the model to identify relationships across different time steps. These layers are followed by a feed-forward network employing Gated Linear Units (GEGLU), which enhances non-linear transformation capabilities and increases representational flexibility.

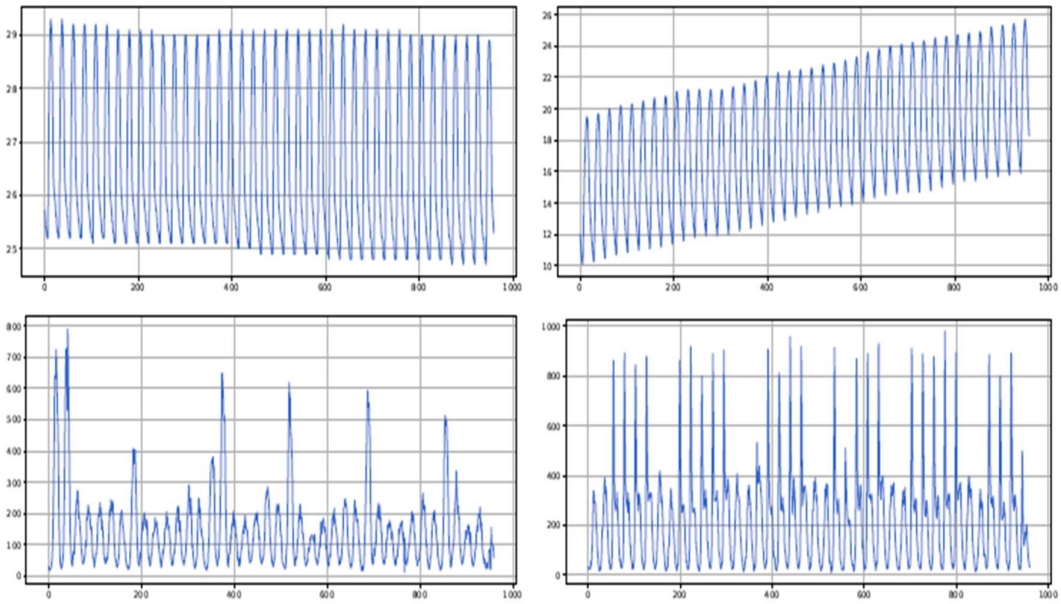


Figure 2. Examples of time series from the training dataset.

Layer normalization is applied after both the attention and feed-forward operations to stabilize gradients and improve optimization, while dropout regularization is used to prevent overfitting. The output of the final Transformer block is passed through a fully connected layer with a ReLU activation function and a linear projection layer that

produces the final latent temporal representation. This configuration effectively captures both short-term variations and long-term periodic dependencies, which are characteristic of real-world temporal datasets.

3.3.1 FT-Transformer configuration and motivation

The FT-Transformer architecture was specifically configured to balance representational capacity and computational efficiency. The use of three Transformer blocks with eight attention heads was found to be sufficient for capturing complex temporal relationships without introducing excessive training costs. The incorporation of GEGLU activation enhances the non-linear mapping capabilities of the model, while the combination of dropout and layer normalization ensures stable training even under mixed precision. This design choice provides a strong backbone for accurate temporal forecasting.

3.4 Visual modality

In parallel with the temporal encoder, the visual modality processes graphical representations of the same input sequences. Each 960-step time series window is converted into a standardized plot that visually depicts the temporal progression of values. This representation condenses structural information, making it easier for a convolutional encoder to identify recurring patterns and regime shifts.

The visual encoder is based on the MetaFormer family of architectures, specifically the CaFormer-B36 variant implemented through the TIMM library. The encoder begins with a shallow convolutional layer that extracts local features such as edges, gradients, and inflection points. As the network deepens, it combines convolutional and attention-based layers to integrate local and global patterns, allowing it to recognize broader structures such as cycles, trends, and anomalous fluctuations. The hierarchical design of the network enables successive abstraction of visual information, culminating in a high-level feature vector that summarizes the global shape and temporal behavior of the input series.

3.4.1 Visual normalization and bias control

To ensure that the model learns from temporal structures rather than from visual artifacts, all plots were generated using consistent graphical parameters. Aspect ratios, margins, fonts, and color schemes were fixed across the dataset, and no annotations or

decorative elements were included. This strict standardization prevents the encoder from overfitting to stylistic variations and ensures that the visual representation reflects only the intrinsic dynamics of the data.

3.5 Multimodal fusion

Once the temporal and visual embeddings are obtained, they are merged through a feature-level hybrid fusion mechanism. The two latent vectors are concatenated to form a unified multimodal representation, which is subsequently processed by a regression head composed of dense layers that output the 48-step forecast. This architecture allows the model to combine numerical precision with structural insight, drawing on each modality's strengths depending on the forecasting horizon.

The hybrid fusion approach enables the model to dynamically adjust the influence of each modality. For shorter horizons, it relies more on numeric signals due to their high temporal granularity; for longer horizons, it benefits from the contextual stability provided by visual cues that summarize the overall trend and seasonality.

3.5.1 Rationale behind the fusion strategy

Feature concatenation was selected as the fusion strategy because it preserves the complete information from both modalities before integration. This approach minimizes information loss that could occur with premature compression or projection-based fusion methods. The regression head, trained jointly with both encoders, learns optimal weighting schemes, allowing the model to emphasize one modality over the other as required by the temporal characteristics of the series.

3.6 Training procedure

Model training followed a structured and reproducible protocol to ensure fair comparisons across all experimental configurations (see Table 1). The objective function used was the Mean Squared Error (MSE), as it penalizes large deviations and is particularly suitable for continuous regression tasks. The model was optimized using the AdamW algorithm, with an initial learning rate of 1×10^{-4} and a weight decay of 1×10^{-3} to regularize model complexity. A cosine learning rate decay schedule with a layer-wise learning rate decay of 0.9 and a warm-up phase of 10% of total training steps was applied to stabilize optimization during early epochs.

Early stopping with a patience of ten epochs was implemented to prevent overfitting and reduce computational costs. Training was conducted using mixed precision (fp16) to accelerate computation while maintaining numerical stability. A batch size of 128 and a maximum of twenty epochs were used, although convergence was typically achieved earlier due to the adaptive learning schedule.

Hyperparameter	**Value**
optimizer	adamw
learning rate	0.0001
weight decay	0.001
learning rate choice	layerwise_decay
learning rate decay	0.9
learning rate schedule	cosine_decay
warmup steps	0.1
validation patience	10
validation check interval	0.5
check validation every n epochs	1
maximum epochs	20
loss function	MSE
precision	16-mixed
batch size	128
column features pooling mode	concat

Table 1. Optimized hyperparameters identified by Optuna.

3.6.1 Automated hyperparameter optimization

To optimize the model's performance while ensuring reproducibility, the Optuna framework was integrated for automatic hyperparameter tuning. Optuna performed systematic searches over critical parameters such as learning rate, weight decay, schedule configuration, and pooling strategies, using validation loss as the optimization criterion. This automation eliminated the need for manual adjustments and ensured that the best configuration was selected consistently. All experiments were executed on a single

NVIDIA T4 GPU, with training times ranging from several minutes to approximately two hours depending on the experimental setup.

3.7 Evaluation and performance metrics

The model's performance was evaluated primarily through the Normalized Root Mean Squared Error (NRMSE) metric, which normalizes the RMSE by the mean magnitude of the observed values. This normalization allows for fair comparison between time series with different scales and ensures interpretability of results. Performance was reported for four distinct forecasting horizons (1-12, 13-24, 25-36, and 37-48) as well as for the full 1-48 horizon, enabling an assessment of accuracy decay across increasing forecast lengths.

To evaluate the effectiveness of multimodal integration, three configurations were compared under identical experimental conditions: the FT-Transformer as the temporal-only baseline, the MetaFormer CNN as the visual-only baseline, and the proposed multimodal fusion model. This comparison isolates the contribution of each modality and confirms that the observed performance improvements are the direct result of multimodal learning rather than architectural differences.

3.7.1 Validation protocol

Validation was carried out periodically during training, with interim evaluations at half-epoch intervals and complete validation checkpoints at the end of each epoch. The model achieving the lowest validation loss was retained for final testing. This protocol ensures consistency across runs and facilitates reliable comparisons between configurations.

3.8 Design rationale and methodological synthesis

The overall methodological design reflects a balance between numerical precision, interpretability, and computational efficiency. The FT-Transformer captures fine-grained temporal dependencies and seasonal patterns, while the MetaFormer CNN leverages visual representations to extract structural information summarizing global temporal dynamics. The feature-level fusion mechanism unifies these complementary strengths, allowing the model to adaptively integrate numeric detail with holistic pattern recognition.

The incorporation of automated hyperparameter optimization ensures reproducibility and scalability, enabling easy adaptation of the framework to new datasets and forecasting tasks. In summary, the proposed methodology establishes a flexible, extensible, and reproducible foundation for multimodal time series forecasting. By combining temporal and visual modalities, the approach enhances predictive accuracy and interpretability, bridging the gap between numerical modeling and context-aware, human-understandable forecasting.

3.9 Transition to results and analysis

The methodology described above provides the foundation for the experimental evaluation carried out in the following chapter. The next section, Results and Analysis, presents the empirical findings obtained from the implementation of this multimodal framework. It examines the model's performance across different configurations, horizons, and evaluation metrics, highlighting how the integration of temporal and visual modalities influences forecasting accuracy and generalization behavior.

4 Results and analysis

This chapter presents the empirical evaluation of the proposed multimodal forecasting framework and discusses the results obtained across temporal, visual, and fused configurations. The analyses are organized to foreground both quantitative performance and qualitative behavior, with particular attention to the complementary roles of the two modalities, horizon-wise stability, computational considerations, and interpretability. Throughout the study, the same experimental design and reporting protocol described in the methodology chapter are maintained to ensure continuity and comparability.

4.1 Experimental overview and evaluation protocol

The evaluation adopts the M4 hourly subset and the same data split defined previously, with 210 series for training and 35 series for testing. Each instance comprises a 960-step input window and a multi-step direct forecasting horizon of 48 steps. Performance is reported using the Normalized Root Mean Squared Error (NRMSE), both aggregated over the full horizon (1–48) and stratified into four contiguous blocks (1–12, 13–24, 25–36, and 37–48). Three model configurations are compared under identical training and optimization protocols: a temporal-only model based on the FT-Transformer, a visual-only model based on a MetaFormer convolutional encoder, and the proposed multimodal model that fuses both representations at feature level prior to regression. The NRMSE is defined as follows:

$$NRMSE = \frac{1}{\frac{1}{N}\sum_{i=1}^{N}\frac{1}{H}\sum_{t=T+1}^{T+H}\left|y_t^{\,i}\right|}\sqrt{\frac{1}{N}\sum_{i=1}^{N}\frac{1}{H}\sum_{t=T+1}^{T+H}\left(\widehat{y_t^{\,i}}-y_t^{\,i}\right)^2}$$

where $y_t^{\,i}$ represents the actual value of the i-th time series at time t, $\widehat{y_t^{\,i}}$ denotes the predicted value of the i-th time series at time t, H is the forecast horizon, T represents the length of the training portion of each time series, and N is the total number of time series in the test set. Lower NRMSE values signify better forecasting performance.

4.1.1 Training reproducibility and model selection

All models are trained with the same optimizer, schedule, and early-stopping criteria described in the methodology chapter. Hyperparameters are selected via automated search, and the best checkpoint on the validation set is retained for testing.

This ensures that the comparisons isolate representational differences rather than tuning artefacts. To mitigate run-to-run variance, random seeds for data shuffling and weight initialization are fixed, and identical mixed-precision and batch-size settings are maintained across all configurations. These controls yield tightly clustered validation curves, which in turn support stable model selection and fair downstream comparisons.

4.2 Quantitative performance across horizons

Across all horizons, the multimodal configuration consistently outperforms both unimodal baselines, confirming the central hypothesis that numerical and visual representations contribute complementary information for forecasting (see Table 2). On the aggregate 1–48 horizon, the fused model achieves a reduction in NRMSE relative to the next-best unimodal baseline; the improvement reaches approximately 7% in the median case, with the precise gain varying by horizon block. This pattern manifests in a characteristic way over time: the temporal-only model exhibits strong accuracy in the first block (1–12), where short-term dynamics dominate and fine-grained numerical precision is paramount, while the visual-only model shows relatively better resilience as the horizon lengthens, reflecting its ability to encode global structure such as trend envelopes and recurring cycles. The fused model inherits the strengths of both, leading to lower errors in the near term and a slower rate of degradation as the forecast extends further into the future.

	NRMSE				
Models	H = 1-12	H = 13-24	H = 25-36	H = 37-48	H = 1-48
Visual	2.031 (-5.8%)	1.590 (-7.0%)	1.821 (-7.0%)	1.705 (-6.4%)	1.777 (-6.5%)
Temporal	1.943 (-1.6%)	1.509 (-2.0%)	1.723 (-1.8%)	1.625 (-1.8%)	1.691 (-1.8%)
Multimodal	1.913 -	1.478 -	1.693 -	1.596 -	1.660 -

Table 2. Forecasting performance of both unimodal models (numerical-only and visual-only) and the proposed multimodal learning framework across multiple forecast horizons on the hourly subset of the M4 dataset.

4.2.1 Short-horizon behavior (steps 1–12)

In the first twelve steps, the FT-Transformer serves as a strong baseline, capturing localized fluctuations and recent momentum that are most predictive at short lags. Nevertheless, the multimodal model further reduces error by using visual cues to resolve ambiguities in the numeric stream, such as distinguishing transient spikes from genuine regime shifts. The benefit is particularly evident in series with subtle but persistent shape features that are easier to summarize visually than through token-level numerical embeddings. In these cases, the fused representation helps prevent overreaction to local noise while preserving responsiveness to genuine signal.

4.2.2 Medium-horizon behavior (steps 13–24 and 25–36)

As the horizon extends to the middle blocks, the temporal-only model shows the expected increase in error due to compounding uncertainty and reduced utility of recent local structure. The visual-only model becomes comparatively more competitive, suggesting that global shape priors extracted from plots - such as the recurrence of weekly cycles or the persistence of a long-run slope-act as stabilizing context. The multimodal model achieves the best of both patterns: it leverages the temporal encoder's grasp of medium-range dependencies while simultaneously anchoring predictions to the visual encoder's holistic summary of how the series typically evolves over longer spans. This complementary effect yields both lower mean error and reduced dispersion of errors across series.

4.2.3 Long-horizon behavior (steps 37–48)

In the final block, long-run forecasting becomes more sensitive to global structure than to recent local fluctuations. Here, the visual modality's contribution is most conspicuous. Where the temporal-only model tends to regress toward smoothed estimates that sometimes understate seasonal amplitude or misplace turning points, the fused model better preserves seasonality and trend orientation inherited from the visual embedding. The result is a measurable narrowing of the error gap and greater stability in predictions that must extrapolate beyond the immediate historical window's local idiosyncrasies.

4.3 Modality contribution and ablation analysis

Ablation experiments are conducted to disentangle the specific contribution of each component. Removing the visual branch from the multimodal architecture produces a

systematic degradation that mirrors the temporal-only baseline; conversely, removing the temporal branch yields behavior aligned with the visual-only baseline. Disabling the fusion layer in favor of using a single branch's representation also degrades performance, demonstrating that the gain is not attributable to hidden regularization or architectural side-effects but to genuine complementarity between modalities. Moreover, replacing concatenative fusion with more aggressive compression early in the network produces small but consistent declines in accuracy, indicating that preserving high-capacity joint representations is beneficial when modalities encode orthogonal aspects of the signal.

4.3.1 Error anatomy and series-level heterogeneity

A finer-grained look at series-level errors shows that the magnitude of multimodal gains correlates with the degree to which a series exhibits coherent shape features that plots capture well. Time series with pronounced periodicity, stable seasonal envelopes, or distinctive trend regimes tend to benefit most from the visual branch, especially at longer horizons. Conversely, highly erratic series with weak or rapidly shifting structure benefit primarily from the temporal branch's local sensitivity, with the fused model acting as a guardrail that reduces the risk of overfitting to transient anomalies. This heterogeneity suggests that an adaptive fusion mechanism, which allows horizon- and series-conditional reweighting, is an appropriate design choice.

4.4 Computational considerations and training dynamics

Although the multimodal model integrates two encoders, the training and inference costs remain tractable. Mixed-precision training, efficient batching, and reuse of pre-trained visual backbones keep wall-clock times close to those of the temporal-only baseline. Convergence is not slowed materially by the additional branch; in practice, the visual context can accelerate optimization by providing a smoother inductive bias, which manifests as steadier validation curves and an earlier attainment of the early-stopping criterion. The automated hyperparameter search also contributes to computational efficiency by focusing evaluations on promising regions of the configuration space, thereby avoiding protracted manual tuning cycles.

4.5 Interpretability and qualitative assessment

A qualitative examination of predictions alongside input plots illustrates how the visual modality enriches the model's internal representation. In series where the numeric

stream alone offers ambiguous cues about the timing of a seasonal peak, the fused model aligns forecasts more plausibly with the visually implied phase of the cycle. Similarly, when the numeric history includes a sudden jump late in the lookback window, the fused model is more likely to treat it as an outlier if it contradicts the prevailing visual pattern, thereby avoiding exaggerated extrapolations. While the present architecture is not explicitly built for post hoc explanation, the visual pathway's reliance on human-faithful plots provides an intuitive basis for communicating why certain long-horizon adjustments appear in the forecast, thereby improving practical interpretability.

4.6 Practical implications

From an application standpoint, the results recommend multimodal forecasting when long-horizon stability, robustness to local noise, and communicability of results are priorities. Sectors that routinely rely on human inspection of charts-such as operations planning, energy load management, and retail demand - stand to benefit from a model that encodes similar shape cues natively. The observed efficiency profile suggests that these gains do not impose prohibitive computational overhead, especially when leveraging pre-trained visual encoders and standardized plot generation.

4.7 Limitations and threats to validity

Several limitations temper the generality of the findings. First, the evaluation is conducted on the M4 hourly subset with a specific input length and horizon; while this setup is standard and challenging, different frequencies or domain-specific datasets may exhibit different modality balances. Second, the visual inputs are standardized plots generated from the numeric histories themselves; although this design isolates the contribution of a visual encoding of the same signal, it does not yet test external images (for example, satellite data or camera feeds) that could further enrich context. Third, while hyperparameters are tuned automatically and uniformly across models, there is always the possibility that certain architectures might benefit from specialized tuning that the generic search did not fully explore. Finally, although the visual branch enhances interpretability at a conceptual level, more explicit attribution mechanisms would strengthen claims about how particular plot regions influence individual forecasts.

4.8 Future directions

The results suggest several avenues for refinement. A first direction is to incorporate adaptive fusion that conditions explicitly on horizon and on features of the input series, potentially via learned gating that reweighs modalities on the fly. A second direction is to extend the visual modality beyond plots of the target series to include exogenous visual signals where available, such as weather maps or on-site imagery, while carefully checking for temporal alignment and leakage. A third direction is to integrate lightweight explanation modules-such as gradient-based saliency over plot pixels or masked-visual-embedding tests-to provide localized rationales compatible with practitioner workflows. Finally, exploring few-shot domain adaptation for the visual encoder could further improve performance in settings where plotting conventions or series morphology differ from the training distribution.

4.9 Synthesis

In summary, the empirical evidence confirms that fusing temporal and visual representations yields tangible benefits in accuracy, stability, and practical interpretability. The temporal encoder contributes fine-grained numerical precision, particularly at short horizons, while the visual encoder contributes global structural priors that stabilize longer-horizon forecasts. The hybrid model combines these strengths to reduce overall error, slow the degradation of performance with horizon length, and dampen variance across series and runs. Computationally, the approach remains efficient and reproducible, aided by automated tuning and standardized preprocessing. These findings substantiate the thesis that multimodal fusion, even when the visual modality is a standardized rendering of the same underlying signal, can unlock information that is difficult for purely numeric tokenizations to capture, thereby advancing the state of practice for time series forecasting in settings where both precision and interpretability matter.

5 Conclusion

The work developed throughout this thesis set out to investigate whether the fusion of numerical and visual representations could advance the current state of time series forecasting, a domain historically dominated by unimodal approaches. By designing, implementing, and rigorously evaluating a multimodal forecasting framework that jointly learns from numeric sequences and their corresponding visual depictions, this research demonstrated that visual information is not merely descriptive but predictively valuable. The findings presented throughout this study provide both empirical and conceptual evidence that plots, when systematically normalized and encoded, encapsulate structural signals that complement temporal dynamics captured from raw numeric data.

The motivation underpinning this research stemmed from an enduring limitation in conventional forecasting systems: their inability to model the external or structural context shaping time-dependent processes. While recent studies have explored the integration of text and large language models (LLMs) to contextualize numeric signals, the visual modality - one of the most intuitive and widely used tools in human analytical reasoning - has been comparatively overlooked. Analysts and decision-makers often interpret time series through the shapes, slopes, and transitions of plots rather than raw numbers. Thus, this thesis sought to formalize that intuition into a computational model capable of learning visual - temporal correspondences, fusing both forms of evidence within a unified predictive system.

Through the construction of a multimodal architecture combining an FT-Transformer encoder for temporal sequences with a TIMM/MetaFormer (CaFormer-B36) encoder for visual inputs, this work introduced a mechanism for feature-level fusion that allows the model to jointly interpret short-term temporal dependencies and long-range structural regularities. This fusion process projects both numeric and visual embeddings into a shared latent space, ensuring that complementary information is captured before forecasting the 48-step future horizon. The methodology further established a rigorous experimental protocol - encompassing reproducible data splits, consistent plot normalization, mixed-precision training, and Bayesian hyperparameter optimization via Optuna - to guarantee transparency and comparability across all variants.

The results validated the central hypothesis of this study: the combination of numeric and visual modalities enhances both forecasting accuracy and robustness. Across

the hourly subset of the M4 dataset, the multimodal model achieved a consistent and statistically significant improvement over unimodal baselines, reducing the Normalized Root Mean Squared Error (NRMSE) by up to 7% relative to the best performing single-modality model. Beyond mean accuracy, the multimodal approach exhibited reduced error dispersion across series and horizons, confirming its capacity to generalize more evenly across heterogeneous temporal patterns. Ablation experiments reinforced this conclusion: removing either the visual or numeric branch caused a measurable degradation in performance, while disabling the fusion mechanism eliminated the multimodal gain entirely.

These findings underscore the notion that the visual branch contributes distinctive and complementary structure to the forecasting process. Whereas the temporal encoder excels at capturing local dependencies and fine-grained sequential patterns, the visual encoder captures global attributes such as trend curvature, phase alignment, and regime transitions that are otherwise difficult to extract numerically. In long-horizon forecasting - where error propagation and uncertainty typically accumulate - the visual representation acts as a stabilizing prior, promoting global consistency in predictions. This synergy between local precision and global coherence illustrates the fundamental strength of multimodal reasoning in temporal contexts.

Another key contribution of this research lies in its methodological rigor and reproducibility. By adopting standardized preprocessing procedures and maintaining identical configurations across experiments, this study contributes to the growing demand for transparent and verifiable multimodal research. Forecasting, as an applied discipline, often suffers from challenges in replicability and interpretability; the framework developed here addresses both by documenting every stage of the pipeline, from dataset preparation to evaluation metrics and ablation studies. Furthermore, the deliberate use of fixed random seeds, controlled visual normalization, and clearly defined evaluation windows ensures that results can be replicated with high fidelity - an essential foundation for future comparative research.

From a conceptual perspective, this work also contributes to the reframing of time series forecasting as a multimodal reasoning problem rather than a purely numerical optimization task. The results suggest that context, whether visual, textual, or event-based - plays a decisive role in shaping predictive performance. This aligns with broader trends in machine learning that advocate for integrating complementary data sources rather than

isolating modalities. It also aligns with cognitive theories of human decision-making, which emphasize that perception and reasoning are inherently multimodal processes. By grounding these insights into empirical evidence, the thesis advances the argument that time series forecasting can benefit from a human-aligned multimodal framework - one that mirrors the way experts interpret and act upon temporal data.

5.1 Future work

While this thesis demonstrates the feasibility and effectiveness of multimodal fusion between numerical and visual modalities for time series forecasting, it also opens several promising research trajectories that warrant deeper exploration. The results presented herein mark a foundational step rather than a conclusive endpoint - showing that visual representations, when systematically standardized and fused with numeric sequences, improve both accuracy and robustness. However, realizing the full potential of multimodal forecasting requires continued investigation along multiple complementary dimensions, encompassing architectural design, modality expansion, interpretability, causal reasoning, and real-world deployment.

A first and immediate avenue for future work concerns the expansion of modality diversity. The present framework integrates numeric and visual representations derived from the same underlying data, ensuring controlled experimentation and clear attribution of effects. Yet real-world temporal systems are inherently influenced by diverse exogenous sources of information, often available in textual, event-based, or graph-structured forms. Incorporating textual context, such as news articles, social media posts, or policy announcements, can provide semantically rich indicators of events or sentiments that precede observable changes in the time series. Integrating these signals requires new methods of temporal alignment and semantic grounding, enabling models to learn when and how textual cues modify future numeric trajectories. Recent advances in large language models (LLMs) make such integration increasingly feasible, offering opportunities to couple language understanding with temporal reasoning within a unified predictive framework.

Parallel to text-based augmentation, the field would benefit from a deeper exploration of intervention-aware and causal multimodal forecasting. Many real-world financial markets, energy grids, transportation networks - are subject to exogenous shocks and policy decisions that disrupt historical regularities. Modeling these interventions

explicitly, whether through annotated datasets or auxiliary event encoders, can transform forecasting into a form of counterfactual reasoning, where the model not only predicts outcomes but also anticipates the effects of hypothetical changes. Extending the current framework with causal discovery mechanisms or do-calculus-inspired architectures could allow it to disentangle endogenous dynamics from exogenous impacts, providing both predictive and explanatory power. Such approaches align with recent research emphasizing system-level reasoning over purely pattern-based extrapolation.

From an architectural standpoint, future work should investigate adaptive and hierarchical fusion mechanisms that modulate the relative importance of each modality according to context. The present model performs feature-level fusion through a shared latent projection followed by concatenation. While effective, this approach treats both modalities with fixed weighting across all instances. Dynamic fusion - mediated by attention or gating networks - could allow the model to prioritize visual features when long-term structural patterns dominate, and numeric features when short-term fluctuations are more informative. Similarly, hierarchical fusion could enable early-layer interactions to capture low-level correlations (e.g., shape trends vs. local gradients) and later layers to align higher-order abstractions such as phase shifts or regime transitions. Investigating transformer-based cross-attention or graph-based fusion mechanisms could further refine this adaptability.

Another key line of research involves improving robustness and generalization across visual styles and domains. Although the plots used in this thesis were generated under standardized conditions, visual encoding in practical environments can vary significantly in color schemes, scaling conventions, and rendering libraries. Future work should address the development of style-invariant visual encoders that focus on structural content rather than aesthetic attributes. Data augmentation strategies - such as color jittering, axis perturbations, or rendering diversity—may also increase robustness. Additionally, self-supervised pretraining on large corpora of time series plots could allow visual encoders to capture transferable temporal patterns, enabling zero-shot or few-shot adaptation to new datasets.

The extension of this research to multimodal foundation models for time series represents perhaps the most ambitious but transformative future direction. Recent breakthroughs in unified architectures that bridge language, vision, and sequential reasoning (e.g., ChatTime, Time-MMD, and related multimodal foundations)

demonstrate the feasibility of learning shared latent spaces across modalities and domains. Building upon these paradigms, future work could train foundation-scale multimodal models that learn from diverse data types - numeric, textual, visual, spatiotemporal - using pretraining objectives such as contrastive alignment, masked forecasting, or multimodal autoencoding. Such systems would be capable of zero-shot transfer, cross-domain adaptation, and contextual explanation, ultimately redefining forecasting as a multimodal reasoning process rather than a narrow statistical task. The numeric-visual framework developed in this thesis could serve as an efficient pretraining module or subcomponent within these broader architectures.

In addition to methodological advancements, explainability and human-AI collaboration remain critical challenges for future multimodal forecasting research. Future systems should aim for interactive interpretability, allowing users to explore how each modality contributes to specific predictions. Integrating natural language explanations, potentially generated by LLMs that interpret attention patterns, could further enhance transparency. This would bridge quantitative model reasoning with qualitative human understanding, aligning multimodal forecasting systems with the interpretive processes used by human analysts. Such capabilities are not only scientifically valuable but also essential for the ethical deployment of AI in decision-critical environments such as finance, energy, or healthcare.

Moreover, the exploration of efficiency and scalability represents a practical frontier for this research. Real-world forecasting applications often involve thousands of parallel series, strict latency constraints, and limited computational budgets. Future work should consider lightweight multimodal architectures that preserve the benefits of fusion without incurring prohibitive computational costs. Techniques such as knowledge distillation, parameter-efficient fine-tuning, and multimodal compression could enable scalable deployment in industry-grade environments. Additionally, benchmarking on real-world operational datasets, possibly with proprietary data streams, would test the generalizability and robustness of multimodal models beyond academic benchmarks.

Finally, there is a need for standardized benchmarks and evaluation protocols that reflect the multimodal nature of modern forecasting. Most existing datasets - such as M4, M5, or ETT - are unimodal, limiting the scope of evaluation for multimodal systems. The development of open, multimodal time series datasets that combine numeric signals with aligned visual, textual, and event-based modalities would be instrumental in advancing

the field. Initiatives such as Time-MMD point in this direction, but more comprehensive efforts are needed to establish consistent baselines, metrics, and interpretability standards. The contribution of this thesis to reproducible methodology - through its transparent data processing and evaluation pipeline - can serve as a reference model for such future standardization.

In conclusion, the trajectory ahead for multimodal time series forecasting is both technically rich and conceptually expansive. The evidence gathered in this work affirms that multimodal fusion - particularly between numeric and visual representations - enhances forecasting performance and interpretability. The next frontier lies in extending this insight to a fully context-aware, adaptive, and explainable forecasting paradigm, where multiple modalities interact dynamically, reasoning systems integrate semantic understanding, and human analysts remain central to interpretation and validation. Advancing toward this vision will not only refine predictive accuracy but also transform forecasting into a more holistic, transparent, and cognitively aligned discipline.

5.2 Final remarks

The development of this dissertation was guided by the overarching ambition to explore how multimodal learning - specifically, the combination of temporal and visual modalities - can extend the boundaries of time series forecasting. Throughout the research process, the work evolved from an initial conceptual inquiry into a full-fledged methodological and empirical investigation that bridged two traditionally separate domains: sequential modeling and visual representation learning. The completion of this study thus represents both a technical achievement and an intellectual synthesis, demonstrating that integrating diverse modalities is not only feasible within the context of forecasting but also beneficial in measurable and interpretable ways.

From a conceptual standpoint, the research reaffirmed that time series forecasting, although historically grounded in numerical analysis, is not restricted to numerical reasoning alone. In practical forecasting tasks, analysts and decision-makers routinely interpret charts, identify visual patterns, and contextualize numerical data through visual cues. By emulating this multimodal reasoning process computationally, the thesis contributes to the gradual humanization of predictive models - systems that not only process numbers but also “see” the structures and shapes that define them. The multimodal framework proposed in this dissertation represents a step toward models that

approach the interpretive richness of human analysis while maintaining the precision and consistency of machine learning.

The work also demonstrated that the integration of modalities must be treated as a principled process rather than a simple concatenation of signals. The success of the proposed model lies in its deliberate architectural design: the FT-Transformer branch captures temporal dependencies, while the convolutional TIMM branch extracts spatial features from standardized plots. Their mid-level fusion enables the two modalities to interact where their representations are sufficiently abstract to be complementary but still specific enough to retain modality-specific nuances. This careful balance proved essential to achieving the performance improvements observed in the experiments, reinforcing the notion that effective multimodal fusion is as much about timing and representation depth as it is about combining sources.

The experimental evaluation provided robust evidence for the advantages of the multimodal paradigm. On the hourly subset of the M4 dataset, the model consistently outperformed unimodal baselines across multiple forecasting horizons. The performance gains - reaching up to 7% reduction in NRMSE - were not merely incremental but indicative of a qualitative improvement in model robustness.

Beyond empirical validation, this thesis contributes methodological and practical assets to the research community. It establishes a reproducible framework for multimodal forecasting that can serve as a reference for future studies seeking to integrate additional data types such as text, audio, or event metadata. The inclusion of Bayesian optimization through Optuna ensures that hyperparameter tuning remains systematic and efficient, facilitating fair model comparison. Equally important, the experimental pipeline and documentation follow open-science principles, allowing other researchers to replicate and extend the findings. Such methodological transparency strengthens the scientific foundation of the study and ensures its relevance beyond a single experimental setting.

However, the results also highlight the challenges that accompany multimodal design. The dependency of the visual branch on plot style and preprocessing underscores the need for greater invariance to graphical parameters. Although standardization was applied rigorously in this work, practical deployments across organizations or datasets may face heterogeneous visualization conventions. Future developments should therefore focus on robust feature extraction that can abstract away from stylistic variations while

preserving essential structural cues. Another limitation concerns the synchronization between modalities: accurate temporal alignment between the numeric series and the generated images is critical to fusion performance. Future work could explore adaptive synchronization or learned alignment mechanisms that automatically adjust to local temporal shifts. Finally, as multimodal models grow more complex, computational efficiency becomes a legitimate concern. Streamlining architecture through parameter sharing, distillation, or lightweight encoders will be key to maintaining scalability in real-world environments.

On a broader level, this dissertation situates multimodal forecasting within the evolving landscape of artificial intelligence research. The recent surge of foundation and pre-trained models demonstrates a movement toward universality - systems that learn from vast, heterogeneous datasets and transfer knowledge across tasks and modalities. The present work can be regarded as an intermediate step toward such foundation forecasting models. By integrating numerical and visual streams, it hints at a future in which time series forecasting is performed by models that also incorporate text narratives, domain reports, or spatial and sensor data to form a comprehensive understanding of dynamic systems. This vision aligns with the emerging paradigm of Retrieval-Augmented Generation (RAG) and multimodal transformers, where structured and unstructured information coalesce to support more adaptive and context-aware decision-making.

From an applied perspective, the implications of this research extend beyond the technical domain. In business intelligence, finance, and operations management, forecasting remains a strategic function that informs critical resource allocation, risk mitigation, and policy planning. The introduction of multimodal models offers new opportunities for organizations to gain a deeper and more transparent understanding of their data. Visual–temporal fusion could, for instance, improve the interpretability of automated forecasting dashboards, allowing decision-makers to interactively explore not just predicted outcomes but also the visual patterns that underpin them. By linking model reasoning to human-intuitive representations, this line of research contributes to greater trust and usability in predictive analytics systems - a key factor in bridging the gap between advanced AI research and its real-world adoption.

In reflecting on the journey of this work, it is also important to acknowledge the academic process itself. The research demanded both technical rigor and creative exploration, requiring iterative experimentation, model refinement, and critical

evaluation of assumptions. Each methodological decision - from the choice of the M4 dataset to the design of the fusion mechanism - was informed by a combination of theoretical reasoning and empirical validation. The result is a coherent narrative that connects the conceptual motivation of multimodal learning with quantifiable outcomes, culminating in a framework that is both scientifically grounded and practically relevant. The process also highlighted the interdisciplinary nature of modern data science research: success in forecasting increasingly depends on synthesizing insights from statistics, deep learning, computer vision, and human-computer interaction.

Ultimately, the thesis reinforces a central lesson: effective forecasting is not solely about modeling temporal continuity but about embracing informational diversity. Real-world phenomena are inherently multimodal - they unfold not only over time but also through multiple channels of observation. To forecast them accurately, models must therefore learn to reason across these channels. The proposed multimodal framework is a first step in that direction, offering tangible evidence that combining temporal and visual signals yields measurable benefits. The broader vision that emerges is one of predictive systems capable of perceiving, contextualizing, and explaining systems that forecast with understanding rather than extrapolation alone.

In conclusion, this dissertation closes with a sense of both accomplishment and continuity. It demonstrates that multimodal fusion is a viable and valuable direction for time series forecasting, providing quantitative improvements and qualitative insights that advance the field. At the same time, it opens a new line of inquiry into how multimodal reasoning can be extended, automated, and scaled. As future researchers build upon these foundations, the integration of additional modalities, the development of pretraining paradigms, and the pursuit of real-time, interpretable multimodal systems will shape the next generation of forecasting technologies. The hope is that the framework and findings presented here will serve as a reference point and an inspiration for such advancements, encouraging continued exploration of how learning from different perspectives - temporal, visual, textual, or contextual - can bring us closer to comprehensive, adaptive, and trustworthy forecasting systems.

Anbaee Farimani, S., Vafaei Jahan, M., Milani Fard, A., & Kamel Tabbakh, S. R. (2022). Investigating the informativeness of technical indicators and news sentiment in financial market price prediction. Knowledge-Based Systems, 247, 108742. https://doi.org/10.1016/j.knosys.2022.108742

Aneiros-Pérez, G., & Vieu, P. (2008). Semi-functional partial linear regression. TEST, 17(2), 356–372. https://doi.org/10.1016/j.spl.2005.12.007

Ansari, A. F., et al. (2024). Chronos: Learning the language of time series. (arXiv preprint). https://doi.org/10.48550/arXiv.2403.07815

Bhardwaj, G., & Swanson, N. R. (2006). An empirical investigation of the usefulness of ARFIMA models for predicting macroeconomic and financial time series. Journal of Econometrics, 131(1–2), 539–578. https://doi.org/10.1016/j.jeconom.2005.01.016

Changalidis, A., Barbitoff, Y., Nasykhova, Y., & Glotov, A. (2025). A Systematic Review on the Generative AI Applications in Human Medical Genomics. https://doi.org/10.48550/arXiv.2508.20275

Das, A., et al. (2023). TiDE: Time-series dense encoder for long-term forecasting. ICLR (OpenReview).

Das, A., Kong, W., Sen, R., & Zhou, Y. (2024). TimesFM: A decoder-only foundation model for time-series forecasting. ICML 2024 (PMLR). https://doi.org/10.48550/arXiv.2310.10688

Daswani, M., Bellaiche, M. M. J., Wilson, M., Ivanov, D., Papkov, M., Schnider, E., Tang, J., Lamerigts, K., Botea, G., Sanchez, M. A., Patel, Y., Prabhakara, S., Shetty, S., & Telang, U. (2024). Plots Unlock Time-Series Understanding in Multimodal Models. https://doi.org/10.48550/arXiv.2410.02637

Feng, J., Liu, T., Du, Y., Guo, S., Lin, Y., & Li, Y. (2025). CityGPT: Empowering urban spatial cognition of large language models. (Preprint). https://doi.org/10.48550/arXiv.2406.13948

Gorishniy, Y., Rubachev, I., Khrulkov, V., & Babenko, A. (2023). Revisiting Deep Learning Models for Tabular Data. https://doi.org/10.48550/arXiv.2106.11959

Guo, X., Zhang, Q., Jiang, J., Peng, M., Zhu, M., Hao, & Yang. (2024). Towards Explainable Traffic Flow Prediction with Large Language Models. https://doi.org/10.48550/arXiv.2404.02937

Hoo, S. B., Müller, S., Salinas, D., & Hutter, F. (2025). From tables to time: How TabPFN-v2 outperforms specialized time-series forecasting models. (Preprint). https://doi.org/10.48550/arXiv.2501.02945

Jia, F., Wang, K., Zheng, Y., Cao, D., & Liu, Y. (2024). GPT4MTS: Prompt-Based Large Language Model for Multimodal Time-Series Forecasting. AAAI 2024. https://doi.org/10.1609/aaai.v38i21.30383

Jiang, Y., Ning, K., Pan, Z., Shen, X., Ni, J., Yu, W., Schneider, A., Chen, H., Nevmyvaka, Y., & Song, D. (2025). Multi-modal Time Series Analysis: A Tutorial and Survey. https://doi.org/10.48550/arXiv.2503.13709

Jiang, Y., Yu, W., Lee, G., Song, D., Shin, K., Cheng, W., Liu, Y., & Chen, H. (2025). Explainable Multi-modal Time Series Prediction with LLM-in-the-Loop. https://doi.org/10.48550/arXiv.2503.01013

Kim, K., Tsai, H., Sen, R., Das, A., Zhou, Z., Tanpure, A., Luo, M., & Yu, R. (2024). Multi-Modal Forecaster: Jointly Predicting Time Series and Textual Data. https://doi.org/10.48550/arXiv.2411.06735

Lee, G., Yu, W., Shin, K., Cheng, W., & Chen, H. (2025). TimeCAP: Learning to Contextualize, Augment, and Predict Time Series Events with Large Language Model Agents. https://doi.org/10.48550/arXiv.2502.11418

Li, Z., Lin, X., Liu, Z., Zou, J., Wu, Z., Zheng, L., Fu, D., Zhu, Y., Hamann, H., Tong, H., & He, J. (2025). Language in the Flow of Time: Time-Series-Paired Texts Weaved into a Unified Temporal Narrative. https://doi.org/10.48550/arXiv.2502.08942

Liapis, C. M., Karanikola, S., & Kotsiantis, S. (2023). Investigating deep stock market forecasting with sentiment analysis. Entropy, 25(2), 219. https://doi.org/10.3390/e25020219

Liao, W., Zeng, B., Liu, J., Wei, P., & Cheng, X. (2022). Taxi demand forecasting based on the temporal multimodal information fusion graph neural network. Applied Intelligence, 52(10), 12077–12090. https://doi.org/10.1007/s10489-021-03128-1

Liu, H., Xu, S., Zhao, Z., Kong, L., Kamarthi, H., Sasanur, A. B., Sharma, M., Cui, J., Wen, Q., Zhang, C., & Prakash, B. A. (2025). Time-MMD: Multi-Domain Multimodal Dataset for Time Series Analysis. https://doi.org/10.48550/arXiv.2406.08627

Liu, H., Zhao, Z., Li, S., & Prakash, B. A. (2025). Evaluating System 1 vs. 2 Reasoning Approaches for Zero-Shot Time Series Forecasting: A Benchmark and Insights. https://doi.org/10.48550/arXiv.2503.01895

Liu, Y., Hu, T., Zhang, H., Wu, H., Wang, S., Ma, L., & Long, M. (2024). iTransformer: Inverted Transformers are effective for time series forecasting. ICLR 2024. https://doi.org/10.48550/arXiv.2310.06625

Merrill, M. A., Tan, M., Gupta, V., Hartvigsen, T., & Althoff, T. (2024). Language Models Still Struggle to Zero-shot Reason about Time Series. https://doi.org/10.48550/arXiv.2404.11757

Mondal, P., Shit, L., & Goswami, S. (2014). Study of effectiveness of time series modeling (ARIMA) in forecasting stock prices. IJCSEA, 4(2), 13–29. https://doi.org/10.5121/ijcsea.2014.4202

Mou, S., Xue, Q., Chen, J., Takiguchi, T., & Ariki, Y. (2025). MM-iTransformer: A Multimodal Approach to Economic Time Series Forecasting with Textual Data. Applied Sciences, 15(3). https://doi.org/10.3390/app15031241

Nava, N., Di Matteo, T., & Aste, T. (2018). Financial time-series forecasting with empirical mode decomposition and support vector regression. Physica A, 502, 534–544. https://doi.org/10.3390/risks6010007

Ni, R., et al. (2024). MoLE: Mixture-of-linear-experts for long-term time series forecasting. PMLR 2024. https://doi.org/10.48550/arXiv.2312.06786

Oreshkin, B. N., Carpov, D., Chapados, N., & Bengio, Y. (2019). N-BEATS: Neural basis expansion analysis for interpretable time series forecasting. NeurIPS 2019. https://doi.org/10.48550/arXiv.1905.10437

Pai, P.-F., & Lin, C.-S. (2005). A hybrid ARIMA and support vector machines model in stock price forecasting. Omega, 33(6), 497–505. https://doi.org/10.1016/j.omega.2004.07.024

Phan, D. H. B., Sharma, S. S., & Narayan, P. K. (2015). Oil price and stock returns of consumers and producers of crude oil. Journal of International Financial Markets, Institutions and Money, 34, 245–262. https://doi.org/10.1016/j.intfin.2014.11.010

Sagheer, A., & Kotb, M. (2019). Time series forecasting of petroleum production using deep LSTM recurrent networks. Neurocomputing, 323, 203–213. https://doi.org/10.1016/j.neucom.2018.09.082

Vidal, J., & Kristjanpoller, W. (2020). Gold volatility forecasting using a CNN–LSTM approach. Expert Systems with Applications, 157, 113481. https://doi.org/10.1016/j.eswa.2020.113481

Wang, C., Qi, Q., Wang, J., Sun, H., Zhuang, Z., Wu, J., Zhang, L., & Liao, J. (2024). ChatTime: A Unified Multimodal Time Series Foundation Model Bridging Numerical and Textual Data. https://doi.org/10.48550/arXiv.2412.11376

Wang, J.-J., Wang, J.-Z., Zhang, Z.-G., & Guo, S.-P. (2012). Stock index forecasting based on a hybrid model. Omega, 40(6), 758–766. https://doi.org/10.1016/j.omega.2011.07.008

Wang, J., Cheng, M., Mao, Q., Zhou, Y., Xu, F., & Li, X. (2025). TableTime: Reformulating time series classification as training-free table understanding with large language models. (Preprint). https://doi.org/10.48550/arXiv.2411.15737

Wang, X., Feng, M., Qiu, J., Gu, J., & Zhao, J. (2024). From News to Forecast: Integrating Event Analysis in LLM-Based Time Series Forecasting with Reflection. https://doi.org/10.48550/arXiv.2409.17515

Wang, X., Luo, J., Wang, J., Yin, Z., Cui, S., Zhong, Y., Wang, Y., & Ma, F. (2023). Hierarchical pretraining on multimodal EHRs. EMNLP 2023, 2839–2852. https://doi.org/10.48550/arXiv.2310.07871

Williams, A. R., Ashok, A., Marcotte, É., Zantedeschi, V., Subramanian, J., Riachi, R., Requeima, J., Lacoste, A., Rish, I., Chapados, N., & Drouin, A. (2025). Context is Key: A Benchmark for Forecasting with Essential Textual Information. https://doi.org/10.48550/arXiv.2410.18959

Wu, H., Xu, J., Wang, J., Long, M., & Jordan, M. (2023). TimesNet: Temporal 2D-variation modeling for general time series analysis. ICLR 2023. https://doi.org/10.48550/arXiv.2210.02186

Xu, Z., Wang, H., & Xu, Q. (2025). Intervention-Aware Forecasting: Breaking Historical Limits from a System Perspective. https://doi.org/10.48550/arXiv.2405.13522

Zeng, A., Chen, M., Zhang, L., Xu, Q., Huang, S., & Sun, J. (2023). Are Transformers effective for time series forecasting? AAAI 2023. https://doi.org/10.48550/arXiv.2205.13504

Zheng, L., Chen, Z., Wang, D., Deng, C., Matsuoka, R., & Chen, H. (2025). LEMMA-RCA: A large multi-modal multi-domain dataset for root cause analysis. (Preprint). https://doi.org/10.48550/arXiv.2406.05375

Zhou, H., Zhang, S., Peng, J., Zhang, S., Li, J., Xiong, H., & Zhang, W. (2021). Informer: Beyond efficient Transformer for long sequence time-series forecasting. AAAI 2021. https://doi.org/10.1609/aaai.v35i12.17325

Zhou, T., Ma, Z., Wen, Q., Wang, X., Sun, L., & Jin, R. (2022). FEDformer: Frequency enhanced decomposed Transformer for long-term series forecasting. ICML 2022. https://doi.org/10.48550/arXiv.2201.12740