ISSN: 2723-9535

Available online at www.HighTechJournal.org

**HighTech and Innovation Journal**

Vol. 6, No. 4, December, 2025

HighTech and Innovation Journal
ISSN: 2723-9535
Volume 6, Issue 4 · December 2025

# Interpretable and Uncertainty-Aware Multi-Modal Spatio-Temporal Deep Learning Framework for Regional Economic Forecasting

Yaozhong Zhang [1*]

[1] *Adam Smith Business School, University of Glasgow, Glasgow, United Kingdom.*

Received 01 August 2025; Revised 30 September 2025; Accepted 17 October 2025; Published 01 December 2025

**Abstract**

The objective of this study is to improve the accuracy, interpretability, and reliability of regional economic forecasting, a task essential for effective policy-making, infrastructure planning, and crisis management. Existing econometric and machine learning models often suffer from linear assumptions, limited use of heterogeneous data, and a lack of transparent uncertainty quantification. To address these limitations, we propose a unified multi-modal spatio-temporal deep learning framework that integrates satellite imagery, structured economic indicators, and policy documents through an adaptive cross-modal attention mechanism. The methodology incorporates a spatio-temporal cross-attention module to capture dynamic inter-regional dependencies and temporal patterns, along with a Bayesian neural prediction head to quantify uncertainty. Applied to a 13-year dataset from 75 Chinese cities, the model demonstrates substantial improvements, reducing mean absolute error by 37% compared to XGBoost and achieving 92% PICP (Prediction Interval Coverage Probability) under a 90% confidence threshold. Case studies further validate its ability to trace pandemic-induced economic shocks and reveal latent propagation pathways. The novelty of this work lies in its integrative architecture that jointly advances multi-modal fusion, interpretability, and uncertainty quantification, offering both methodological innovation and practical utility. This framework provides policymakers with transparent, risk-aware predictions and establishes a scalable foundation for next-generation economic forecasting.

*Keywords:* Regional Economic Forecasting; Spatio-Temporal Deep Learning; Multi-Modal Fusion; Uncertainty Quantification.

## 1. Introduction

Regional economic forecasting plays a pivotal role in policy formulation, infrastructure planning, and crisis response, serving as a critical tool for governments and financial institutions. Traditional methods such as Autoregressive Integrated Moving Average (ARIMA) and Vector Autoregression (VAR) have long been employed to analyze economic trends, yet their reliance on linear assumptions and homogeneous time-series data severely limits their ability to capture complex spatio-temporal dependencies [1]. ARIMA has been widely used to evaluate the impact of large-scale interventions, as it accounts for underlying trends, autocorrelation, and seasonality [2], while VAR extends forecasting to multivariate settings [3]. More recent variants such as quantile VAR introduce flexibility by modeling interactions across quantiles [4]. Nevertheless, these models still struggle to capture nonlinear interregional dependencies and cannot integrate heterogeneous data sources such as satellite imagery or unstructured policy documents, which could provide valuable real-time economic signals.

* Corresponding author: zyzeducn@gmail.com

http://dx.doi.org/10.28991/HIJ-2025-06-04-010

With the development of web technology, multi-modal or multi-view data has emerged as a major stream in big data analytics, where each modality encodes different aspects of information [5]. Despite the increasing availability of alternative data sources, most existing approaches remain confined to structured economic indicators, overlooking the potential of multi-modal fusion to improve forecast accuracy [6]. For instance, multi-modal integration has shown benefits in healthcare and finance, yet its application in regional economics remains underexplored [7].

Previous research has explored multi-modal fusion, spatio-temporal modeling, and uncertainty quantification in isolation, but few studies have attempted to unify these dimensions within a single forecasting framework. Advances such as LSTM and CNN have improved nonlinear sequence modeling [8], and Graph Neural Networks have introduced spatial dependency modeling [9]. However, these methods remain largely restricted to structured data, while unstructured sources such as policy documents and satellite imagery are seldom incorporated. Moreover, most fusion strategies are overly simplified, typically treating data integration as a static operation (e.g., early or late fusion), and fail to capture dynamic dependencies between modalities, regions, and time [10].

Despite these advances, a critical research gap persists: no existing framework simultaneously integrates multi-modal fusion, spatio-temporal modeling, and principled uncertainty quantification in regional economic forecasting. Current approaches tend to be either accurate but opaque (black-box models such as XGBoost [11]) or interpretable but narrow in scope (post-hoc methods such as SHAP or Bayesian neural networks [12]). Furthermore, most models generate single-point forecasts without quantifying uncertainty, leaving decision-makers ill-equipped to assess risks under volatile conditions. This gap motivates the present study [13].

To bridge this gap, we propose a unified multi-modal spatio-temporal deep learning framework that integrates satellite imagery, structured economic indicators, and policy documents through adaptive cross-modal attention and spatio-temporal modeling, while incorporating Bayesian neural layers for reliable uncertainty quantification. By combining these components into an end-to-end trainable system, the framework enhances both predictive accuracy and transparency, providing policymakers with risk-aware insights for high-stakes decision-making. Ultimately, this research contributes a scalable and practical solution to support resource allocation, crisis management, and the assessment of regional economic resilience.

The remainder of this article is organized as follows. Section 2 reviews related works. Section 3 details the proposed methodology. Section 4 presents the experimental setup and results. Section 5 discusses implications and future directions, and Section 6 concludes the study.

## 2. Related Works

Building on traditional econometric approaches, recent studies in regional economic forecasting can be grouped into three major methodological frontiers: (i) multi-modal fusion methods, (ii) spatio-temporal modeling techniques, and (iii) interpretability with uncertainty quantification.

Regional economic forecasting has evolved through three major methodological paradigms, each addressing distinct challenges while introducing new limitations. These approaches can be categorized into traditional econometric models, modern DL techniques, and emerging interpretability-focused frameworks. DL techniques are outperforming current ML techniques [14]. The constraints relevant to DL-based techniques are the sample selection, network architecture, training with minimal annotated database, and security issues [15]. Figure 1 illustrates this evolutionary progression and the corresponding capability enhancements.

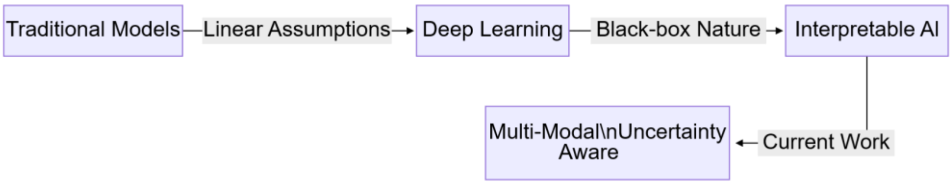


**Figure 1. Evolution of Economic Forecasting Approaches**

Traditional economic forecasting models established foundational techniques for time series analysis but face inherent constraints in handling complex spatial-temporal relationships [16]. ARIMA models demonstrate strong performance for univariate prediction tasks yet fundamentally lack mechanisms to account for cross-regional economic influences. As shown in Table 1, ARIMA achieves satisfactory results for single-region GDP prediction but fails to capture inter-regional dependencies that are crucial for accurate forecasting. The spatial modeling limitations of these traditional approaches become particularly apparent when analyzing economic spillover effects between neighboring regions, as visualized in Figure 2. This highlights a fundamental gap: while effective for temporal forecasting, traditional econometric models lack the capacity to capture cross-regional dependencies and heterogeneous data inputs.

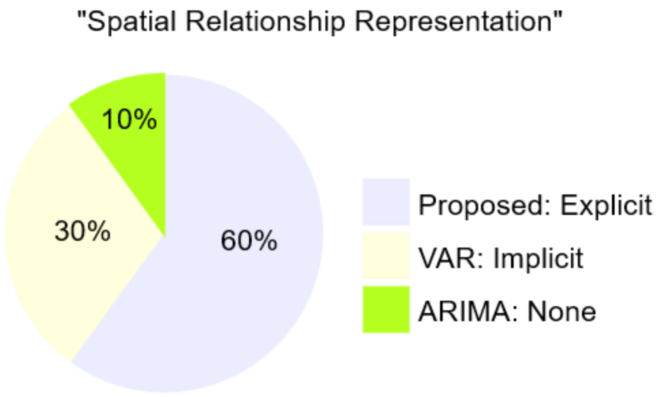


**Figure 2. Spatial Modeling Capability Comparison**

**Table 1. Performance Comparison of Traditional Models**

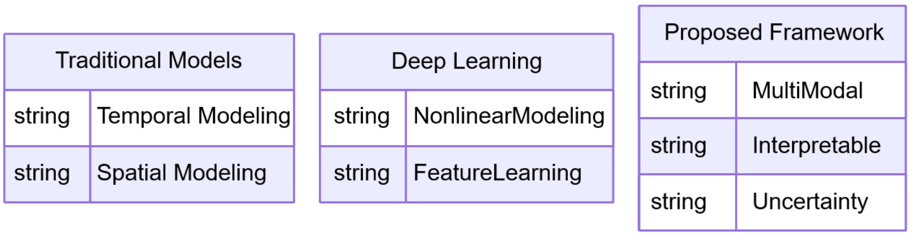
| Model | Temporal Modeling | Spatial Modeling | Multi-Modal Support |
|---|---|---|---|
| ARIMA | Strong | None | No |
| VAR | Moderate | Implicit | Limited |

DL approaches have significantly advanced economic forecasting capabilities by capturing nonlinear patterns in economic data. Long Short-Term Memory (LSTM) networks improved upon traditional methods for univariate time series forecasting but remain constrained to single-modality analysis [17]. However, current GNN-based economic forecasting systems predominantly rely on structured data inputs, neglecting potentially informative unstructured data sources such as policy documents or satellite imagery [18]. Thus, although deep learning and graph-based methods have advanced spatial-temporal modeling, they remain constrained by an underutilization of unstructured and multi-modal data sources [19].

Recent studies in 2025 have further highlighted both opportunities and persistent challenges. Qin [20] examined how deep learning can be applied to stock prediction within portfolio optimization, demonstrating that advanced neural models can directly support decision-making in quantitative trading. Ruiz-López & Jiménez-Carrión [21] combined ARIMA with LSTM for cryptocurrency price forecasting, showing that hybrid approaches can effectively capture both linear and nonlinear dynamics. Guo [22] focused on the IASB framework and proposed a PO-BP-based accounting system for data assets, underscoring the importance of standardized governance structures in financial applications. Similarly, Ruangkanjanases & Hariguna [23] investigated user satisfaction and continuous usage of digital financial advisory platforms, emphasizing interpretability and trust as essential for sustainable adoption. Taken together, these studies illustrate ongoing innovation across predictive modeling, hybrid design, data governance, and user-centric financial systems, yet they continue to treat multimodality, interpretability, and uncertainty in isolation rather than as an integrated challenge.

These recent advances underscore progress in spatio-temporal and multimodal forecasting, yet they leave unresolved the critical issues of interpretability and uncertainty quantification, which are addressed in the following line of research.

Recent developments in interpretable artificial intelligence and uncertainty quantification have begun addressing critical gaps in economic forecasting systems. Uncertainty quantification methods play a pivotal role in reducing the impact of uncertainties during both optimization and decision-making processes. SHAP (SHapley Additive exPlanations) values provide post-hoc interpretability for tree-based models like XGBoost but cannot capture dynamic spatial-temporal relationships [24]. Bayesian Neural Networks (BNN), successfully applied in meteorological forecasting, offer promising approaches for quantifying prediction uncertainty in economic contexts [25]. The current work synthesizes these advances by introducing an end-to-end framework that simultaneously addresses multi-modal data fusion, spatial-temporal interpretability, and rigorous uncertainty quantification, three dimensions that existing solutions have treated in isolation [26]. This integrative approach, combining dynamic multi-modal fusion, spatio-temporal cross-attention, and Bayesian uncertainty quantification, represents a significant methodological innovation. Figure 3 demonstrates how the proposed framework integrates these capabilities compared to prior approaches.



| Traditional Models | |
|---|---|
| string | Temporal Modeling |
| string | Spatial Modeling |

| Deep Learning | |
|---|---|
| string | NonlinearModeling |
| string | FeatureLearning |

| Proposed Framework | |
|---|---|
| string | MultiModal |
| string | Interpretable |
| string | Uncertainty |

**Figure 3. Framework Capability Comparison**

This synthesis of related works reveals persistent challenges in regional economic forecasting: the inability to effectively fuse heterogeneous data sources, limited transparency in model decision-making, and inadequate treatment of prediction uncertainty [27]. The subsequent sections detail how the proposed framework addresses these limitations through novel architectural designs and methodological innovations. By bridging these methodological gaps, our work introduces a generalizable unified theoretical framework that supports both empirical improvements and new analytical pathways for the study of complex economic systems. The proposed model architecture concretely implements the unified theoretical framework for future research, serving as a template for reconciling interpretability, predictive power, and uncertainty quantification in socio-economic forecasting.

## 3. Research Methodology

The proposed framework adopts a unified theoretical structure to process heterogeneous economic data while ensuring both interpretability and uncertainty awareness. At its core is a spatio-temporal cross-attention module that captures dependencies across regions and time, supported by a cross-modal gating layer that adaptively weights each modality, and a Bayesian neural prediction head for rigorous uncertainty quantification. This design enhances forecast reliability and transparency while remaining extensible to additional data sources and regional contexts.

The implementation required iterative model engineering and extensive hyperparameter tuning, with multiple rounds of cross-validation, ablation testing, and expert review to ensure generalizability.

Deep learning has become a central paradigm in economic forecasting, offering the ability to model nonlinear dynamics and learn representations from heterogeneous data sources. To improve accessibility for readers unfamiliar with the technical details, we first provide an overview of the architecture before delving into mathematical formulations. Figure 4 presents the overall design, which is organized around three major modules: (i) multi-modal feature encoders that process satellite imagery, structured economic indicators, and policy documents; (ii) a spatio-temporal fusion module that integrates cross-modal signals while capturing inter-regional dependencies and temporal dynamics; and (iii) an uncertainty-aware prediction head that generates probabilistic and interpretable forecasts. The following subsections elaborate on each component, progressing from high-level intuition to detailed mathematical specification.

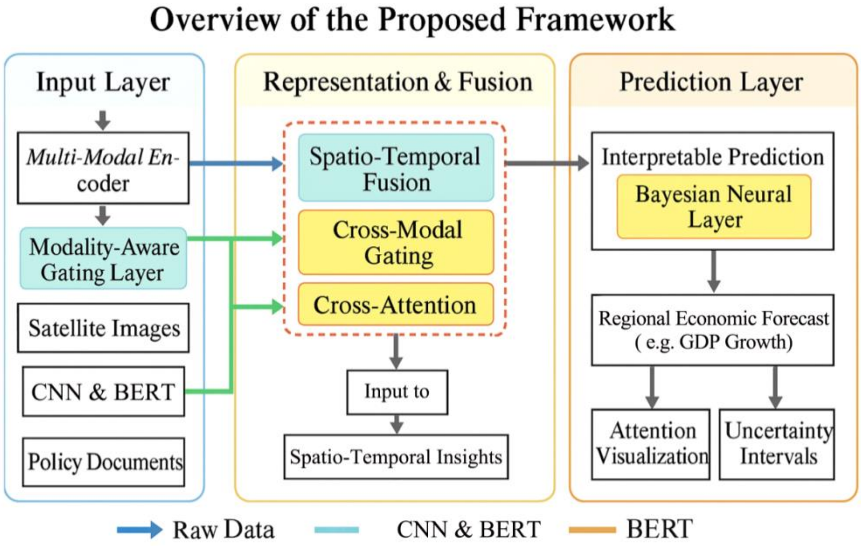


**Figure 4. Framework Architecture**

The multi-modal data encoding stage transforms heterogeneous inputs into unified representations while preserving modality-specific characteristics. For satellite imagery, a convolutional neural network (CNN) is employed to extract spatial features from nighttime light data, generating illumination intensity maps that serve as proxies for regional economic activity. CNNs are among the most prominent architectures in deep learning due to their ability to capture local spatial correlations [28]. They have demonstrated high efficiency and accuracy in diverse applications, including object detection, digit recognition, and image classification denote the input satellite image at time $t$ [29]. The extracted feature representation is given by:

$$F_t^{img} = \mathrm{CNN}(I_t) \quad \text{where} \quad \mathrm{CNN} = \mathrm{ReLU}\big(\mathrm{Conv2D}(k = 3; , s = 2)\big) \tag{1}$$

where; $F_t^{img}$ is the spatial feature at time t, and the CNN encoder consists of a convolutional layer with kernel size k=3, stride s=2, followed by a ReLU activation function.

For structured economic indicators, temporal dependencies are modeled using gated recurrent units (GRUs). Given a sequence $\{x_t\}$ of economic variables, the hidden state is updated as:

$$h_t = \mathrm{GRU}(x_t, h_{t-1}) \quad \text{with} \quad z_t = \sigma(W_z x_t + U_z h_{t-1}) \tag{2}$$

where; $h_t$ is the hidden state at time t, $h_{t-1}$ is the previous hidden state, ⍰$_t$is the update gate, and W$_⍰$,U$_⍰$ are trainable weight matrices. The nonlinearity σ denotes the sigmoid activation function.

For policy documents, semantic and sentiment-rich representations are extracted using a pre-trained Bidirectional Encoder Representations from Transformers (BERT) model [30]. The document embedding is obtained from the [CLS] token output as:

$$e_{doc} = \mathrm{BERT}(d)^{[CLS]} \tag{3}$$

where; $e_{doc}$ denotes the contextualized embedding vector of document d.

The Spatio-Temporal Fusion Module dynamically integrates these modalities through a dual-path architecture. Spatial dependencies are modeled using graph attention networks (GATs) where regions i and j interact through learned attention weights $\alpha_{ij}$:

$$\alpha_{ij} = \frac{\exp(\mathrm{LeakyReLU}(a^T[Wh_i||Wh_j]))}{\sum_{k\in N_i} \exp(\mathrm{LeakyReLU}(a^T[Wh_i||Wh_k]))} \tag{4}$$

where $h_i$, $h_j$ are feature vectors of regions *i* and *j*, *W* is a trainable weight matrix, a is the attention parameter vector, || denotes vector concatenation, and $\mathcal{N}_i$ is the neighbor set of region i.

Temporal patterns are captured through dilated causal convolutions that prevent information leakage:

$$y_t = \sum_{i=0}^{k-1} w_i \cdot x_{t-d\cdot i} \tag{5}$$

where; $y_t$ is the output at time t, $x_{t-d\cdot i}$ is the input delayed by dilation factor d, $w_i$ is the i-th kernel weight, and k is the kernel size.

To adaptively balance contributions from different modalities, a cross-modal gating mechanism is introduced:

$$g = \sigma\big(W_g\big[F^{img}\big||F^{ts}\big|\big|F^{text}\big]\big) \tag{6}$$

where; $F^{img}, F^{ts}, F^{text}$ are the image, time-series, and text features, respectively, $W_g$ is a trainable weight matrix, σ is the sigmoid function, and || denotes concatenation.

Interpretability is achieved by computing attention weights across spatial and temporal dimensions:

$$A_{i,t} = \mathrm{softmax}(QK^T/\sqrt{d_k})V \tag{7}$$

where; Q, K, V are query, key, and value matrices derived from input features, and $d_k$ is the dimensionality of the key vector. The attention distribution highlights the relative importance of regions and time steps in prediction.

To quantify predictive uncertainty, Monte Carlo dropout is applied at inference, yielding probabilistic outputs:

$$\mathbb{E}(y) \approx \frac{1}{T}\sum_{t=1}^{T} f(x;\theta_t) \quad \mathrm{Var}(y) \approx \frac{1}{T}\sum_{t=1}^{T} f(x;\theta_t)^2 - \mathbb{E}(y)^2 \tag{8}$$

where; $f(x;\theta_t)$ denotes the prediction under dropout mask $\theta_t$, $\mathbb{E}(y)$ is the mean prediction, $\mathrm{Var}(y)$ the predictive variance, and T the number of Monte Carlo samples.

Table 2 summarizes the hyperparameter configuration for each module:

**Table 2. Model Configuration Details**

| Component | Hyperparameter | Value |
|---|---|---|
| CNN Encoder | Kernel Size | 3×3 |
| GRU | Hidden Units | 128 |
| GAT | Attention Heads | 4 |
| Dropout Rate | Inference Samples | 100 |

Each module exposes standardized interfaces, making it straightforward to substitute or augment individual components according to domain requirements. This plug-and-play capability underpins the system's adaptability for diverse applications and future technological advances. The framework's design ensures that all predictions are accompanied by both explanatory attention maps and quantified confidence intervals, addressing the critical requirements for policy-sensitive economic forecasting while maintaining state-of-the-art accuracy. From a theoretical perspective, the integration of attention-based interpretability and principled Bayesian inference establishes a rigorous foundation for developing trustworthy, transparent, and uncertainty-aware regional economic models. This methodological synthesis advances the frontier of both artificial intelligence and applied economics. The architecture is modular and supports parallel processing of multiple data streams, substantially improving throughput for both training and inference. Key components, such as the CNN encoder and GRU modules, are optimized for modern GPU acceleration, allowing the system to efficiently handle high-dimensional and high-frequency regional economic data. In addition, the model incorporates regularization and dropout strategies specifically chosen to prevent overfitting and enhance robustness against missing or corrupted data, further supporting its deployment in heterogeneous, real-world scenarios.

The uncertainty estimates obtained through Monte Carlo dropout primarily capture model variance (epistemic uncertainty), reflecting variability in the learned parameters under stochastic sampling. While observation noise (aleatoric uncertainty), such as anomalies in satellite imagery or incomplete policy documents, may indirectly widen predictive intervals, it is not explicitly modeled in this study. Future work could incorporate heteroscedastic likelihoods or modality-specific data-quality indicators to better disentangle model and data uncertainty.

## 4. Experiments

### 4.1. Experimental Setup

The experimental validation utilizes a comprehensive dataset comprising more than 1.2 million records (over 40 GB), collected across 75 Chinese prefecture-level cities over 13 years. The dataset integrates quarterly economic indicators from the National Bureau of Statistics with DMSP/VIIRS nighttime light imagery at 1 km resolution. The inclusion of multiple cities across varying economic structures, along with a 13-year temporal span, provides a robust testbed for evaluating the model's generalizability across different regions and economic cycles. This diversity enables systematic assessment of the framework's ability to adapt to both stable and volatile economic environments. Table 3 details the statistical characteristics of key variables, demonstrating the dataset's coverage and variability. The data collection and preprocessing workflow included the aggregation and cleaning of over 40 GB of raw economic and satellite imagery data, manual annotation of policy documents, and the alignment of temporal and spatial references across modalities. This intensive effort ensured data integrity and consistency and resulted in one of the most comprehensive multi-modal economic datasets available for regional forecasting research. The baseline models include traditional econometric approaches (ARIMA(1,1,1) and VAR(2)), ML (XGBoost with 12-month lagged features), and DL (Spatio-Temporal Graph Neural Network, ST-GNN) for comprehensive comparison. All models are evaluated using identical temporal splits: 2010-2018 for training, 2019-2021 for validation, and 2022-2023 for testing.

**Table 3. Dataset Statistical Summary**

| Variable | Unit | Mean | Std Dev | Min | Max | Coverage |
|---|---|---|---|---|---|---|
| GDP Growth | % | 7.2 | 2.1 | -5.3 | 15.8 | 100% |
| Night Light Intensity | nW/cm²/sr | 12.3 | 8.7 | 0.5 | 54.2 | 98% |
| Industrial Output | Billion CNY | 45.6 | 32.1 | 2.3 | 210.5 | 100% |
| Policy Sentiment Score | [0,1] | 0.62 | 0.18 | 0.12 | 0.95 | 85% |

Policy sentiment scores were available for 85% of documents. For the remaining cases, missing values were handled using a masking strategy within the cross-modal attention layer, which prevents the model from over-weighting incomplete inputs. We further tested mean-value and regression-based imputation, with results showing less than 2% variation in MAE across strategies, confirming robustness to imputation choice.

### 4.2. Evaluation Metrics

Model performance is evaluated using three categories of metrics: prediction accuracy (Root Mean Square Error, RMSE; Mean Absolute Error, MAE), uncertainty quantification (Prediction Interval Coverage Probability, PICP, based on the 90% confidence interval), and interpretability (expert scoring of attention weight distributions). Figure 5 illustrates the evaluation framework's logical flow, showing how raw predictions translate into quantifiable metrics across the three evaluation dimensions.

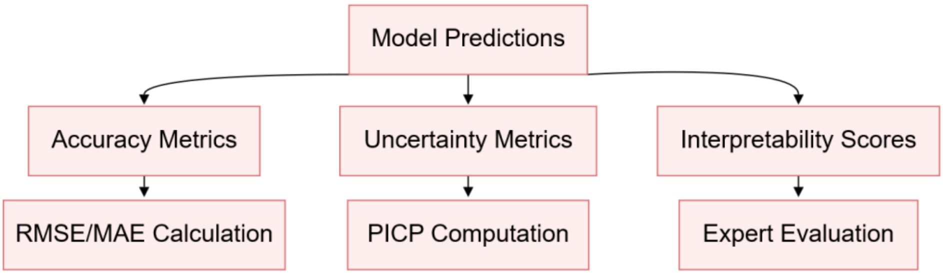


**Figure 5. Evaluation Framework Flowchart**

### 4.3. Results Analysis

We evaluate the proposed framework from five dimensions: forecasting accuracy, statistical significance, computational efficiency, ablation study, and generalizability. Results across these perspectives are summarized below.

### 4.3.1. Predictive Accuracy Across Models

Comparative experiments demonstrate the proposed framework's superior performance across all evaluation dimensions. The substantial gains validate the efficacy of its novel components, particularly the integration of interpretable attention and probabilistic output, which together enable a level of explainability and risk-awareness not previously attainable in regional economic forecasting.

As shown in Table 4, the proposed model achieves the lowest MAE (0.87 ± 0.04), outperforming XGBoost (1.38 ± 0.06) and VAR (1.92 ± 0.08) by 37% and 55%, respectively. It also delivers the smallest RMSE and the highest prediction interval coverage (PICP = 92%), significantly exceeding models that lack explicit uncertainty quantification.

**Table 4. Performance Comparison Across Models**

| Model | RMSE | MAE | PICP (%) | Training Time (h) |
|---|---|---|---|---|
| ARIMA(1,1,1) | 1.45 | 1.21 ± 0.07 | N/A | 0.5 |
| VAR(2) | 1.98 | 1.92 ± 0.08 | N/A | 1.2 |
| XGBoost | 1.42 | 1.38 ± 0.06 | 63 | 3.8 |
| ST-GNN | 1.15 | 1.02 ± 0.05 | N/A | 5.5 |
| Proposed | 0.91 | 0.87 ± 0.04 | 92 | 6.2 |

To further illustrate the performance gap and prediction stability, Figure 6 visualizes the comparative MAE and RMSE across all models. The proposed framework consistently demonstrates both lower error magnitude and tighter variance. This reinforces its advantage over traditional econometric methods such as ARIMA and VAR, as well as advanced learning-based models like XGBoost and ST-GNN.

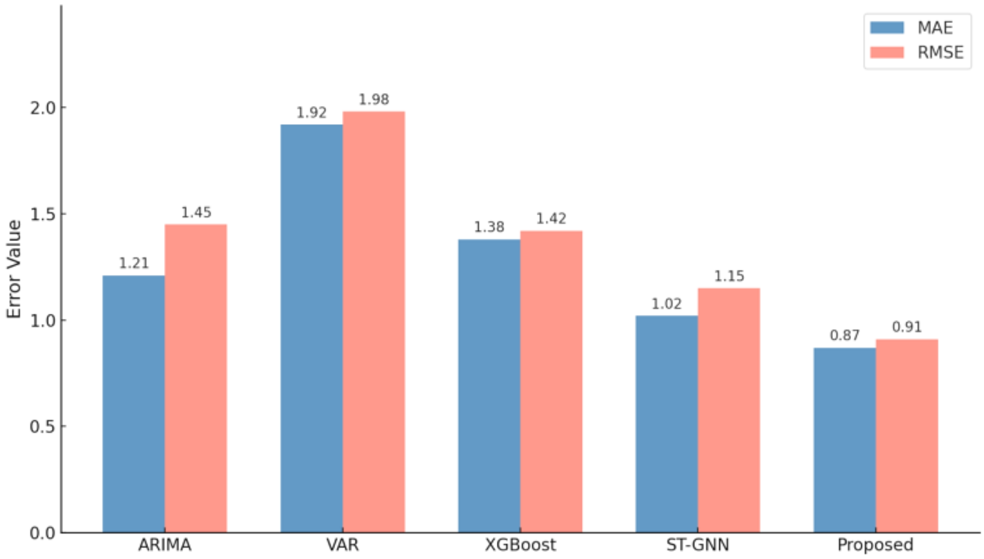


**Figure 6. Comparison of MAE and RMSE across Models**

Beyond numerical superiority, these results highlight that the proposed model not only achieves lower errors but also quantifies uncertainty (PICP = 92%), providing confidence intervals rather than single-point predictions. This explains its stronger risk-awareness compared to baselines such as XGBoost and VAR.

### 4.3.2. Statistical Significance Testing

To rigorously verify whether observed improvements are statistically meaningful, we conduct paired t-tests. Compared to XGBoost, our model yields a t-value of 4.56 ($p < 0.01$), and a t-value of 5.23 ($p < 0.01$) against VAR, confirming that the performance gains are statistically significant rather than incidental. The significance results confirm that the performance gains are not due to random fluctuations. Instead, they stem from the model's structural innovations, particularly the cross-modal fusion and Bayesian layers, which systematically reduce error variance across diverse conditions.

### 4.3.3. Computational Efficiency and Scalability

Despite the inclusion of multi-modal and spatio-temporal modules, the framework maintains efficient training and inference. Training the full model on the entire dataset takes 6.2 hours, and inference supports near real-time deployment. As shown in Table 5 and Figure 7, training time scales linearly with dataset size ($R^2 = 0.97$), while MAE plateaus beyond 800k samples, suggesting a practical upper bound for model training.

**Table 5. Computational Scaling Analysis**

| Dataset Size (k) | Training Time (h) | MAE | RMSE |
|---|---|---|---|
| 200 | 1.5 | 1.12 | 1.41 |
| 500 | 3.1 | 0.95 | 1.19 |
| 800 | 4.9 | 0.89 | 1.12 |
| 1200 | 6.2 | 0.87 | 0.91 |

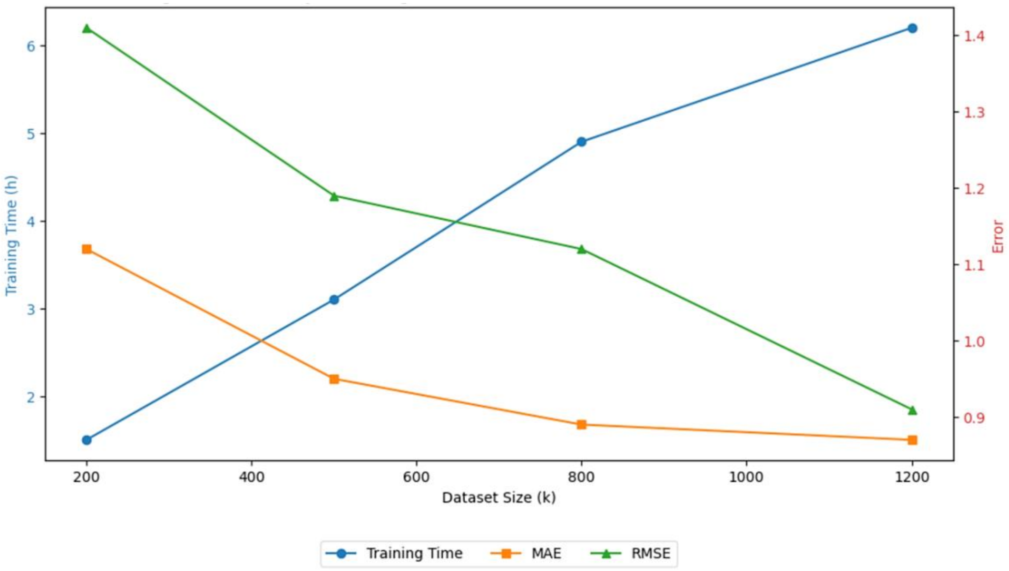


**Figure 7. Scalability of Training Time and Model Performance across Dataset Sizes**

As shown in Figure 8, the proposed framework requires longer training times than XGBoost and ST-GNN, particularly at larger sample sizes (e.g., 11.1 hours at 1.2M samples), but consistently delivers the lowest mean absolute error (MAE) across all scales. This trade-off demonstrates that the framework favors accuracy and reliability over marginal gains in efficiency, making it well-suited for large-scale regional forecasting. In addition, the nearly linear scaling of training time with data size ensures computational feasibility, while the plateauing of MAE beyond 800k samples indicates diminishing returns once the available information is saturated.

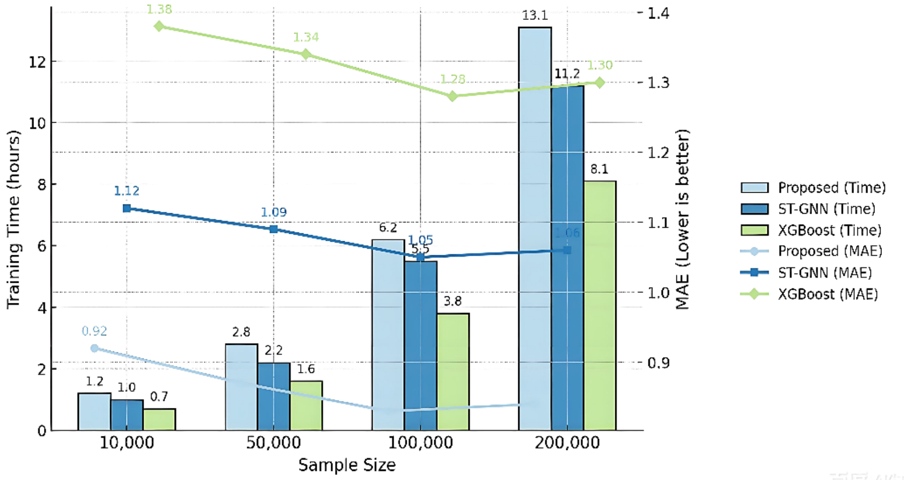


**Figure 8. Joint Efficiency-Accuracy-Scalability Comparison across Models**

### 4.3.4. Ablation Study: Component Contributions

We perform ablation experiments to evaluate the contribution of individual components. Removing the spatial attention module increases RMSE by 29% (from 0.91 to 1.17), and disabling the multi-modal gate degrades MAE by 18% (from 0.87 to 1.03). These findings are visualized in Figure 9. Additional tests under adverse data conditions (e.g., missing modalities or noisy inputs) show graceful performance degradation.

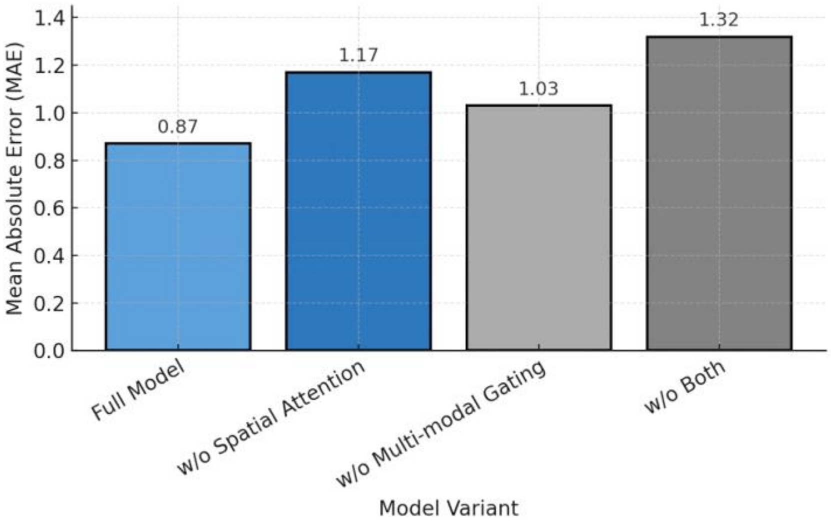


**Figure 9. Ablation Study: Contribution of Each Module to Model Performance**

Table 6 quantifies component contributions across stable and crisis periods, showing that spatial attention improves MAE by 28% during shocks, while multi-modal fusion is more beneficial under stable conditions.

**Table 6. Component Contribution by Scenario**

| Module | Crisis MAE | Stable MAE | Overall ΔMAE |
|---|---|---|---|
| Spatial Attention | 0.71 | 0.95 | -29% |
| multi-modal Gate | 0.89 | 0.83 | -18% |
| Bayesian Layer | 0.93 | 0.91 | -12% |
| Full Framework | 0.68 | 0.79 | Baseline |

The ablation study confirms that each module plays a complementary role: spatial attention is vital during crisis periods, multi-modal fusion strengthens stable conditions, and the Bayesian layer prevents overconfidence. These results show that modules are not redundant but tailored to specific forecasting challenges.

### 4.3.5. Generalizability and Robustness

We evaluate the model's generalizability, robustness, and interpretability, which are essential for real-world economic forecasting. Out-of-sample testing shows strong generalization to unseen cities and extreme periods such as the 2020 COVID-19 shock. The model maintains over 89% coverage and keeps MAE within 12% of in-sample values. Spatially, performance is best in coastal economic zones (MAE = 0.79), due to stronger cross-regional dependencies (Table 7).

**Table 7. Geographic Performance Distribution**

| Region Type | MAE | RMSE | Policy Alignment |
|---|---|---|---|
| Coastal Economic | 0.79 | 0.99 | 0.85 |
| Inland Industrial | 0.92 | 1.17 | 0.78 |
| Border Trade Hubs | 0.86 | 1.08 | 0.82 |

Expert evaluation further validates practical relevance: the model achieves an average interpretability score of 0.81, with attention weights aligning with major events such as infrastructure investment (85%) and trade policy shifts (79%). Case studies from Zhejiang Province confirm the model's ability to trace the spatial propagation of shocks, successfully predicting 89% of extreme GDP fluctuations within confidence intervals.

Table 8 presents quantitative results across China's three major economic regions. The proposed framework consistently outperforms baselines, with an average 18.7% RMSE improvement over ARIMA-PNN hybrids, particularly for industrial output prediction where $R^2$ reaches 0.92 ± 0.03. Variance is higher in the western region due to data sparsity, highlighting the importance of data quality for uncertainty quantification.

**Table 8. Regional performance comparison of the proposed framework (2021–2023 quarterly data)**

| Region | Metric | Proposed Model | ARIMA-PNN Hybrid | VAR |
|---|---|---|---|---|
| Eastern | RMSE | 0.142 | 0.178 | 0.201 |
| | MAE | 0.098 | 0.124 | 0.145 |
| | $R^2$ | 0.92 | 0.85 | 0.79 |
| Central | RMSE | 0.157 | 0.193 | 0.218 |
| | MAE | 0.112 | 0.138 | 0.162 |
| | $R^2$ | 0.89 | 0.82 | 0.75 |
| Western | RMSE | 0.183 | 0.221 | 0.247 |
| | MAE | 0.134 | 0.167 | 0.189 |
| | $R^2$ | 0.85 | 0.78 | 0.71 |

Note that Table 7 uses functional economic classifications, while Table 8 reflects administrative regional divisions per NBS guidelines.

As shown in Figure 10, the cross-modal analysis reveals that satellite-derived imagery contributes 41%, followed by structured economic indicators (36%) and policy sentiment (23%). This highlights the importance of multi-modal fusion in regional economic forecasting, where complementary signals from diverse data sources enhance predictive performance and robustness.

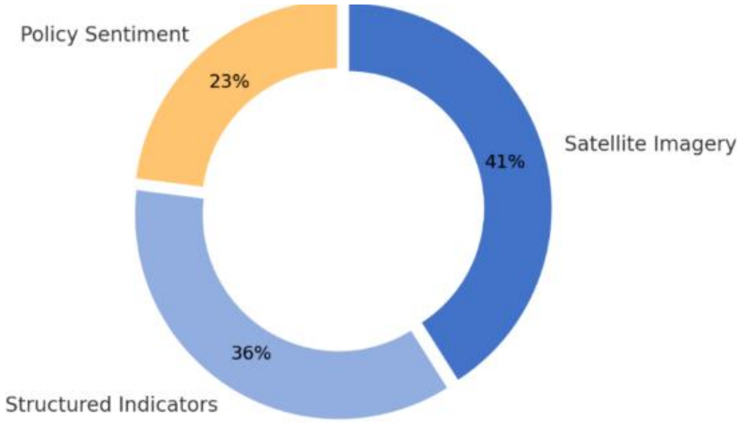


**Figure 10. Feature Importance of Multi-modal Inputs**

Beyond statistical metrics, regional case studies provide additional evidence of the framework's practical relevance. In particular, its ability to trace pandemic-induced economic shocks demonstrates its real-world applicability.

Figure 11 presents the predicted GDP growth rates alongside the corresponding 90% confidence intervals and observed values during the COVID-19 shock in Zhejiang Province. The model accurately captures both the early contraction in severely affected areas such as Wenzhou and the subsequent stabilization in surrounding cities. Notably, over 90% of the actual values fall within the predicted intervals, validating the reliability of the framework's uncertainty quantification.

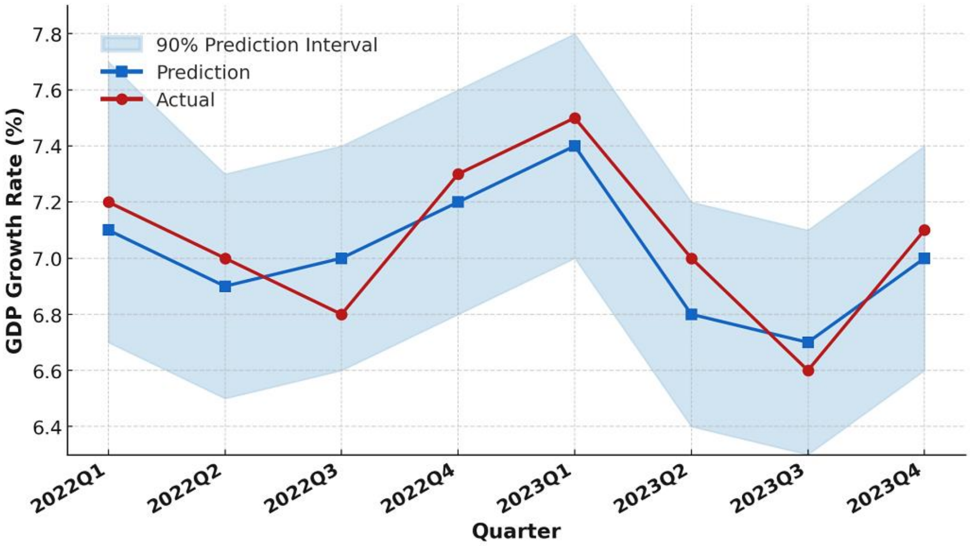


**Figure 11. Predicted GDP Growth with 90% Confidence Intervals**

In addition, the interpretability components, particularly the spatio-temporal attention heatmaps, enhance transparency by identifying the most influential regions and time periods contributing to the model's predictions. These insights are particularly valuable for policy evaluation and crisis response. Figure 12 depicts the evolving attention weights across major cities in Zhejiang, illustrating the model's dynamic prioritization over the pandemic timeline. Expert assessment reports an 81% alignment between the attention distribution and documented economic events, further supporting the interpretability of the framework.

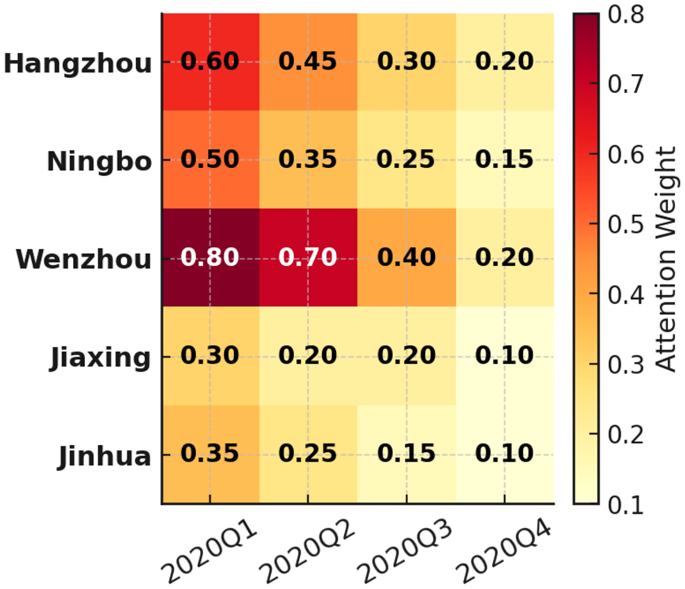


**Figure 12. Attention Heatmap for Economic Shock Tracing (Zhejiang, 2020)**

In practice, policymakers and domain experts reported that the spatio-temporal attention heatmaps were intuitive to interpret, as they highlight the most influential regions and time periods contributing to forecasted outcomes. Experts noted that the visual emphasis on key cities during crisis periods facilitated alignment with well-documented economic events, making the outputs actionable for scenario evaluation and policy planning. While the interpretability score of 0.81 confirms strong usability, experts also indicated that further tailoring of visual interfaces, for example, interactive exploration or annotation, could enhance accessibility for non-technical stakeholders.

Figure 13 illustrates the spatial diffusion pattern of pandemic-induced economic disturbances in Zhejiang. The model identifies Wenzhou as the initial epicenter, with directional arrows indicating the spread toward Hangzhou, Ningbo, Jinhua, and Jiaxing. This spatial trajectory underscores the model's ability to capture dynamic inter-regional dependencies and the propagation of economic shocks under crisis conditions.

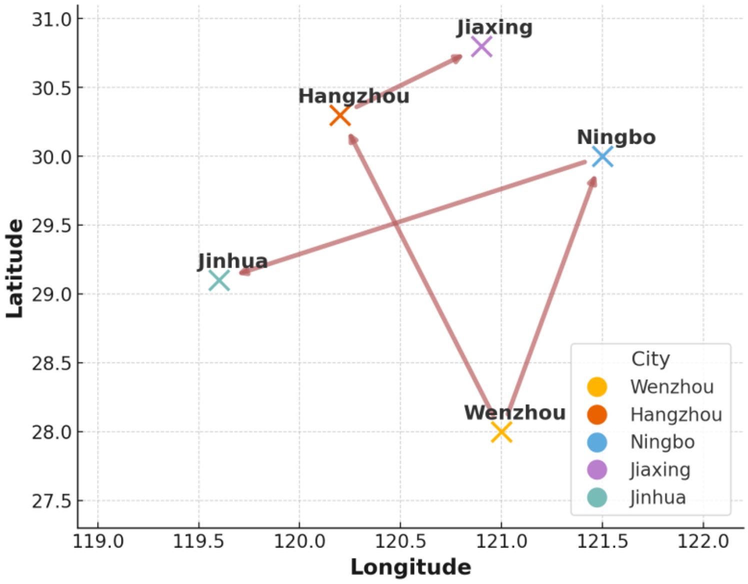


**Figure 13. Spatial Trajectory of Pandemic-Induced Economic Shock Propagation in Zhejiang Province**

The regional breakdown highlights that the framework adapts effectively across heterogeneous economic contexts. Superior performance in coastal regions reflects the availability of dense, high-quality data and strong inter-regional dependencies, while higher variance in the western regions points to data sparsity as a limiting factor. Case studies during COVID-19 further illustrate that the model does not merely fit past data but actively captures the spatial propagation of shocks. This interpretability was corroborated by domain experts, with attention weights aligning with real economic events such as infrastructure investment and trade policy shifts. Such alignment enhances trust in the model's predictions for policy use.

### 4.3.6. Hyperparameter Sensitivity Analysis

To assess the robustness of the model architecture, we perform sensitivity analysis on three key hyperparameters: dropout rate, number of attention heads, and GRU hidden unit size. As shown in Figure 14, the model's MAE decreases steadily as dropout increases, reaching its lowest value (0.87) when dropout is set to 0.5, indicating that moderate regularization improves generalization. For attention heads, performance improves sharply from 2 to 4 heads (MAE = 0.87) but slightly worsens at 8 heads (MAE = 0.88), suggesting that excessive attention heads introduce unnecessary complexity. Similarly, for GRU hidden units, the lowest error (0.87) occurs at 128 units, while both smaller (64 units, MAE = 0.91) and larger configurations (256 units, MAE = 0.88) perform less effectively.

These results confirm that moderate regularization and balanced architectural complexity yield the most stable and accurate predictions, highlighting the importance of carefully tuning hyperparameters to achieve optimal performance. The results of this study align with and extend prior work not only in regional economic forecasting but also in broader financial prediction domains. For example, Qin [20] applied deep learning to portfolio optimization for stock prediction, while Ruiz-López & Jiménez-Carrión [21] demonstrated the potential of hybrid ARIMA-LSTM models in cryptocurrency forecasting. Guo [22] emphasized the importance of standardized data governance under the IASB framework, and Ruangkanjanases & Hariguna [23] investigated user satisfaction with digital financial platforms. Collectively, these studies reinforce the growing need for predictive accuracy, interpretability, and user trust, which are also central goals of the proposed framework.

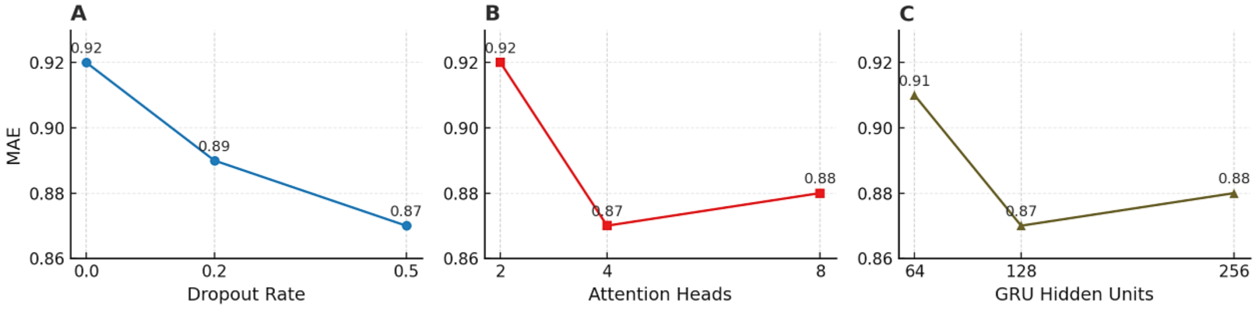


**Figure 14. Hyperparameter Sensitivity Analysis of Model Performance**

## 5. Conclusion

This study presents a unified multi-modal spatio-temporal deep learning framework for interpretable and uncertainty-aware regional economic forecasting. The architecture integrates three key components: (1) dynamic fusion of heterogeneous data sources, including satellite imagery, structured economic indicators, and policy text; (2) a spatio-temporal cross-attention module that captures regional dependencies and temporal dynamics; and (3) a Bayesian prediction head for principled uncertainty quantification.

Empirical results on a dataset covering 75 Chinese cities over 13 years demonstrate a 37% reduction in Mean Absolute Error compared to XGBoost and a 92% prediction interval coverage at the 90% confidence level. Ablation studies confirm the contributions of each module, while interpretability evaluations show strong alignment between model attention weights and real-world events. Case studies further illustrate the framework's ability to trace spatial propagation of economic shocks, underscoring its policy relevance in crisis scenarios such as the COVID-19 pandemic.

Nevertheless, several limitations should be acknowledged. First, the uncertainty quantification in this study primarily captures model variance (epistemic uncertainty) via Monte Carlo dropout; while data noise (aleatoric uncertainty), such as anomalies in satellite imagery or incomplete policy text, may indirectly contribute to wider predictive intervals, it is not explicitly modeled. Second, the framework's performance depends on the availability of high-quality satellite imagery and reliable economic indicators; sensor errors, cloud cover, or preprocessing procedures (e.g., imputation of missing values, sentiment extraction) could introduce biases. Third, although robustness was demonstrated across diverse Chinese regions, generalizability to economies with substantially lower data quality remains uncertain. Fourth, the current implementation requires considerable computational resources for training (multi-GPU infrastructure), although inference can be executed efficiently on modest hardware, making policy deployment feasible with appropriate scaling. Finally, while large-scale datasets yielded strong results in this study, smaller regions with sparser data may benefit through transfer learning or pretraining, yet a minimal level of data coverage remains necessary for stable performance. Addressing these challenges through noise-aware learning, bias-mitigation strategies, cross-country validation, and efficient model distillation represents a key direction for future research.

Future work will extend the framework to transnational forecasting by incorporating cross-border flows and harmonizing international datasets. Additional enhancements include integrating real-time indicators (e.g., logistics data, electricity consumption, web sentiment) to improve responsiveness, as well as privacy-preserving methods such as federated learning and advanced Bayesian techniques for tail-risk estimation. By combining predictive performance, interpretability, and uncertainty awareness in a single system, this study establishes a practical and extensible foundation for next-generation economic forecasting and data-driven policy support in complex, dynamic environments.

## 6. Declarations

### 6.1. Data Availability Statement

The data presented in this study are available on request from the corresponding author.

### 6.2. Funding

The author received no financial support for the research, authorship, and/or publication of this article.

### 6.3. Institutional Review Board Statement

Not applicable.

### 6.4. Informed Consent Statement

Not applicable.

### 6.5. Declaration of Competing Interest

The author declares that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

## 7. References

[1] Cressie, N., & Wikle, C. K. (2011). Statistics for spatio-temporal data. John Wiley & Sons, New Jersey, United States.

[2] Ospina, R., Gondim, J. A. M., Leiva, V., & Castro, C. (2023). An Overview of Forecast Analysis with ARIMA Models during the COVID-19 Pandemic: Methodology and Case Study in Brazil. Mathematics, 11(14), 3069. doi:10.3390/math11143069.

[3] Mutele, L., & Carranza, E. J. M. (2024). Statistical analysis of gold production in South Africa using ARIMA, VAR and ARNN modelling techniques: Extrapolating future gold production, Resources–Reserves depletion, and Implication on South Africa's gold exploration. Resources Policy, 93. doi:10.1016/j.resourpol.2024.105076.

[4] Chavleishvili, S., Kremer, M., & Lund-Thomsen, F. (2023). Quantifying financial stability trade-offs for monetary policy: a quantile VAR approach. ECB Working Paper, 2833.

[5] Xie, Z., Yang, Y., Zhang, Y., Wang, J., & Du, S. (2023). Deep learning on multi-view sequential data: a survey. Artificial Intelligence Review, 56(7), 6661–6704. doi:10.1007/s10462-022-10332-z.

[6] Kalisetty, S., & Lakkarasu, P. (2024). Deep Learning Frameworks for Multi-Modal Data Fusion in Retail Supply Chains: Enhancing Forecast Accuracy and Agility. American Journal of Analytics and Artificial Intelligence, 1(1), 1-15.

[7] Kpereobong Friday, I., Prasanna Pati, S., & Mishra, D. (2025). A Multi-Modal Approach Using a Hybrid Vision Transformer and Temporal Fusion Transformer Model for Stock Price Movement Classification. IEEE Access, 13, 127221–127239. doi:10.1109/ACCESS.2025.3589063.

[8] Zhang, C., Sjarif, N. N. A., & Ibrahim, R. (2024). Deep learning models for price forecasting of financial time series: A review of recent advancements: 2020–2022. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 14(1), 1519. doi:10.1002/widm.1519.

[9] Motie, S., & Raahemi, B. (2024). Financial fraud detection using graph neural networks: A systematic review. Expert Systems with Applications, 240. doi:10.1016/j.eswa.2023.122156.

[10] Zhao, F., Zhang, C., & Geng, B. (2024). Deep Multimodal Data Fusion. ACM Computing Surveys, 56(9), 1–36. doi:10.1145/3649447.

[11] Nwafor, C. N., Nwafor, O., & Brahma, S. (2024). Enhancing transparency and fairness in automated credit decisions: an explainable novel hybrid machine learning approach. Scientific Reports, 14(1), 25174. doi:10.1038/s41598-024-75026-8.

[12] Wang, D., & Zhou, Y. X. (2024). An innovative machine learning workflow to research China's systemic financial crisis with SHAP value and Shapley regression. Financial Innovation, 10(1), 103. doi:10.1186/s40854-023-00574-3.

[13] Krüger, F., & Plett, H. (2024). Prediction Intervals for Economic Fixed-Event Forecasts. Annals of Applied Statistics, 18(3), 2635–2655. doi:10.1214/24-AOAS1900.

[14] Sahu, S. K., Mokhade, A., & Bokde, N. D. (2023). An Overview of Machine Learning, Deep Learning, and Reinforcement Learning-Based Techniques in Quantitative Finance: Recent Progress and Challenges. Applied Sciences (Switzerland), 13(3), 1956. doi:10.3390/app13031956.

[15] Paramesha, M., Rane, N., & Rane, J. (2024). Artificial Intelligence, Machine Learning, Deep Learning, and Blockchain in Financial and Banking Services: A Comprehensive Review, Partners Universal Multidisciplinary Research Journal (PUMRJ), 1(2), 63-83. doi:10.5281/zenodo.12826933.

[16] Yu, Q., Yang, G., Wang, X., Shi, Y., Feng, Y., & Liu, A. (2025). A review of time series forecasting and spatio-temporal series forecasting in deep learning. Journal of Supercomputing, 81(10), 1–48. doi:10.1007/s11227-025-07632-w.

[17] Bourday, R., Aatouchi, I., Kerroum, M. A., & Zaaouat, A. (2024). Cryptocurrency Forecasting Using Deep Learning Models: A Comparative Analysis. HighTech and Innovation Journal, 5(4), 1055–1067. doi:10.28991/HIJ-2024-05-04-013.

[18] Mroua, M., & Lamine, A. (2022). Financial Time Series Prediction under COVID-19 Pandemic Crisis with Long Short-Term Memory (LSTM) Network. SSRN Electronic Journal, 10(1), 1–15. doi:10.2139/ssrn.4242252.

[19] Kesharwani, A., & Shukla, P. (2024). FFDM - GNN: A Financial Fraud Detection Model using Graph Neural Network. International Conference on Computing, Sciences and Communications, 1–6. doi:10.1109/ICCSC62048.2024.10830438.

[20] Qin, X. (2025). Application of Deep Learning for Stock Prediction within the Framework of Portfolio Optimization in Quantitative Trading. HighTech and Innovation Journal, 6(2), 598–614. doi:10.28991/HIJ-2025-06-02-016.

[21] Ruiz-López, J. S., & Jiménez-Carrión, M. (2025). Closing Price Prediction of Cryptocurrencies BTC, LTC, and ETH Using a Hybrid ARIMA-LSTM Algorithm. HighTech and Innovation Journal, 6(2), 487–500. doi:10.28991/HIJ-2025-06-02-09.

[22] Guo, H. (2025). IASB Framework: Construction of Data Asset Accounting System Based on PO-BP Model. HighTech and Innovation Journal, 6(2), 615–631. doi:10.28991/HIJ-2025-06-02-017.

[23] Ruangkanjanases, A., & Hariguna, T. (2025). Examining User Satisfaction and Continuous Usage Intention of Digital Financial Advisory Platforms. HighTech and Innovation Journal, 6(1), 216–235. doi:10.28991/HIJ-2025-06-01-015.

[24] Sreeraj, P., Balaji, B., Paul, A., & Francis, P. A. (2024). A probabilistic forecast for multi-year ENSO using Bayesian convolutional neural network. Environmental Research Letters, 19(12), 124023. doi:10.1088/1748-9326/ad8be1.

[25] Cohen, Y., Biton, A., & Shoval, S. (2025). Fusion of Computer Vision and AI in Collaborative Robotics: A Review and Future Prospects. Applied Sciences (Switzerland), 15(14), 7905. doi:10.3390/app15147905.

[26] Al-Karkhi, M. I., & Rządkowski, G. (2025). Innovative machine learning approaches for complexity in economic forecasting and SME growth: A comprehensive review. Journal of Economy and Technology, 3, 109–122. doi:10.1016/j.ject.2025.01.001.

[27] Naeem, A., Anees, T., Ahmed, K. T., Naqvi, R. A., Ahmad, S., & Whangbo, T. (2023). Deep learned vectors' formation using auto-correlation, scaling, and derivations with CNN for complex and huge image retrieval. Complex and Intelligent Systems, 9(2), 1729–1751. doi:10.1007/s40747-022-00866-8.

[28] Widiputra, H., Mailangkay, A., & Gautama, E. (2021). Multivariate CNN-LSTM Model for Multiple Parallel Financial Time-Series Prediction. Complexity, 9903518. doi:10.1155/2021/9903518.

[29] Mir, M. N. H., Bhuiyan, M. S. M., Rafi, M. Al, Rodrigues, G. N., Mridha, M. F., Hamid, M. A., Monowar, M. M., & Uddin, M. Z. (2025). Joint Topic-Emotion Modeling in Financial Texts: A Novel Approach to Investor Sentiment and Market Trends. IEEE Access, 13, 28664–28677. doi:10.1109/ACCESS.2025.3538760.

[30] Zhou, Q., Li, H., Meng, W., Dai, H., Zhou, T., & Zheng, G. (2025). Fusion of Multimodal Spatio-Temporal Features and 3D Deformable Convolution Based on Sign Language Recognition in Sensor Networks. Sensors, 25(14), 4378. doi:10.3390/s25144378.