

Big Data Applications and Intelligent Decision Support Framework for Study Tourism Market Demand Forecasting

Ping Yan

yanpingchina@outlook.com

Shanxi Vocational University of Cultural Tourism

Research Article

Keywords: Study tourism demand forecasting, big data fusion, cross-modal attention mechanism, Transformer-LSTM, intelligent decision support framework

Posted Date: October 29th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11157298/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: No competing interests reported.

Big Data Applications and Intelligent Decision Support Framework for Study Tourism Market Demand Forecasting

Received:

Accepted:

published online:

Cite this article as:

Big Data Applications and Intelligent Decision Support Framework for Study Tourism Market Demand Forecasting

Ping Yan^{1,a}

¹Shanxi Vocational University of Cultural Tourism, Tai yuan ,Shanxi, China

^aE-mail: yanpingchina@outlook.com

Abstract

This paper addresses three critical bottlenecks in study tourism market demand forecasting: data heterogeneity, time series non-stationarity, and decision-making lag, proposing a big data-driven intelligent decision support framework termed BDIDS (Big Data Intelligent Decision Support). The framework comprises a three-tier architecture consisting of a data fusion layer, a predictive computation layer, and a decision output layer. The data layer integrates multi-modal sources including travel platforms, social media, and policy documents through standardized feature engineering pipelines fitted exclusively on training-period observations to prevent leakage. The predictive layer introduces a cross-modal attention fusion mechanism, coordinating BERT semantic embeddings with time series statistical features to drive an improved Transformer-LSTM hybrid model. The decision layer embeds demand elasticity analysis and supply-demand matching optimization modules. Using a provincial-monthly panel dataset from four East China provinces/municipalities spanning 2018–2024, 336 total province-month cells are constructed; after excluding 24 cells corresponding to complete operational suspension during COVID-19 control periods, 312 valid province-month observations remain, partitioned into 216 training, 48 validation, and 48 test observations ($216+48+48=312$), with the full panel index provided as supplementary material for independent verification. On the 48-observation 2024 test set, BDIDS achieves RMSE reductions versus ARIMA and single LSTM under a pre-specified Diebold–Mariano protocol (MSE loss, Newey–West HAC lag=4, Bonferroni-corrected across seven comparisons); results are conditional on training-period specification and constitute within-sample evidence. Accuracy metrics in ten thousand person-trips support industry resource allocation decisions.

Keywords: Study tourism demand forecasting; big data fusion; cross-modal attention mechanism; Transformer-LSTM; intelligent decision support framework

1. Introduction

As an emerging business form merging education and tourism, study tourism has demonstrated consistent annual market growth exceeding 15% since policy promotion in 2017, with increasingly complex demand structures. Market demand is driven nonlinearly by multiple factors including seasonal cycles, educational policies, and unexpected public events, causing traditional ARIMA-type time series models to experience significant accuracy degradation when facing high-dimensional heterogeneous driving factors [1]. Online booking data, social media sentiment, and policy documents constitute massive multi-modal information sources; how to efficiently fuse and convert these into interpretable prediction signals remains an important scientific question [2]. Recent advances in Transformer architecture [3] and pre-trained language models such as BERT [4] have demonstrated meaningful progress in time series forecasting and semantic understanding, providing new technical pathways for multi-source data fusion modeling [5]. However, existing research generally remains at the single prediction model level, lacking integrated frameworks that combine data fusion, end-to-end forecasting, and actionable decision outputs with auditable evaluation evidence [6].

This paper makes three primary, independently testable contributions: (1) a multi-modal data fusion pipeline for study tourism demand forecasting, built on a documented provincial-monthly panel with explicit exclusion rules and a panel index enabling independent sample verification; (2) an improved Transformer-LSTM hybrid model with cross-modal gated attention and a learnable seasonal

bias matrix, validated against seven baseline methods under a pre-specified and Bonferroni-corrected comparison protocol, with results reported as conditional within-sample evidence; and (3) a decision layer with supply-demand matching optimization, whose marginal contribution relative to prediction improvement alone is evaluated via a historical holdout backtest under identical capacity and cost assumptions, with uncertainty explicitly reported. Each contribution addresses a gap relative to existing work: (1) extends data pipeline design relative to single-source approaches [2,6]; (2) integrates domain-adapted seasonal bias into the attention mechanism, building on standard Transformer baselines [3,5]; (3) embeds supply-demand matching into a forecasting framework, extending prediction-only pipelines [11,12]. The BDIDS framework forms a verifiable technical pipeline from data collection to decision output, holding practical significance for enhancing operational capabilities in the study tourism market.

2. Study Tourism Demand Characteristics and Big Data Collection Architecture

2.1 Demand-Driving Factors and Multi-Source Data Association Modeling

Study tourism demand is driven jointly by three primary factors: educational policy shocks, school holiday scheduling, and destination resource endowments, forming a compound time series structure combining strong seasonality with non-stationary trends [7]. Descriptive analysis on the 216-observation training sample (January 2018–December 2022, after excluding 24 COVID-related suspension cells) reveals that policy document sentiment polarity changes lead market demand fluctuations by approximately 2–4 weeks. Online search indices and actual booking volumes exhibit a Pearson correlation coefficient of approximately 0.73 in the training period. Social media topic heat peaks demonstrate associations with team registration volumes that are statistically distinguishable from zero in Granger causality tests on the training sample ($p < 0.05$). These statistics are reported as descriptive measures of the training data and should not be interpreted as causal identification. Further introducing VAR models for dynamic simultaneous modeling of three categories of driving factors, impulse response function analysis suggests that policy shock effects estimated on the training sample peak around lag week 3 before gradually attenuating, providing econometric guidance for feature time-window specification. A multi-factor association driving matrix is constructed, jointly encoding policy intensity, sentiment heat, and time series statistical characteristics into a multi-modal driving tensor with dimensions $\mathbb{R}^{(T \times (ds+dl))}$, where time series dimension T and lag order are determined by AIC optimization criteria. VIF multicollinearity diagnostics are performed on the three categories of driving factors using the training period, with results showing all factors have VIF below 3.2, indicating limited multicollinearity among the specified driving pathways in this dataset, establishing data-semantic foundations for subsequent cross-modal fusion modeling.

2.2 Multi-Modal Data Collection, Cleaning, and Feature Engineering Pipeline

Multi-modal data sources encompass structured traffic data from online booking platforms (Ctrip, Meituan), unstructured text sentiment from social platforms (Weibo, Xiaohongshu), and policy documents released by the Ministry of Education and Department of Culture and Tourism [8] (see Figure 1). Online booking traffic data and search index data are obtained respectively through platform data-sharing cooperation agreements (signed 2021) and Baidu Index academic authorization APIs (API version accessed March 2021–January 2025); social media texts are collected through each platform's open APIs in strict compliance with each platform's developer terms of service; only aggregated, anonymized content is retained with no personally identifiable information stored or processed; policy documents derive from publicly released government websites, accessed between January 2018 and January 2025. All data collection and processing activities involve publicly available platform data and government documents only, with no human subjects research, no collection of individual personal information, and no experimental intervention on participants. This study was confirmed by the

institutional review authority of Shanxi Vocational University of Cultural Tourism as not requiring ethics committee approval, on the grounds that no human subjects research, clinical intervention, or collection of identifiable personal data is conducted; the institutional confirmation is available upon request. All data usage adheres to applicable data protection regulations and platform terms. For data that cannot be fully released publicly, a controlled-access audit arrangement can be made available to reviewers upon reasonable request, with the specific contact path detailed in the Data Availability statement. The collection pipeline implements real-time streaming data ingestion based on Apache Kafka, with offline batch data entering distributed feature storage following ETL cleaning. The feature engineering pipeline comprises three serial steps executed in strict training-only-fit order: missing value time series interpolation (cubic spline method, fitted on 216 training-period observations only), anomaly detection (Isolation Forest, trained on the training set), and standardization/normalization (mean and variance computed on the training set and applied without re-fitting to validation and test sets). This pipeline ultimately outputs unified-dimensional multi-modal feature tensors for predictive layer utilization.

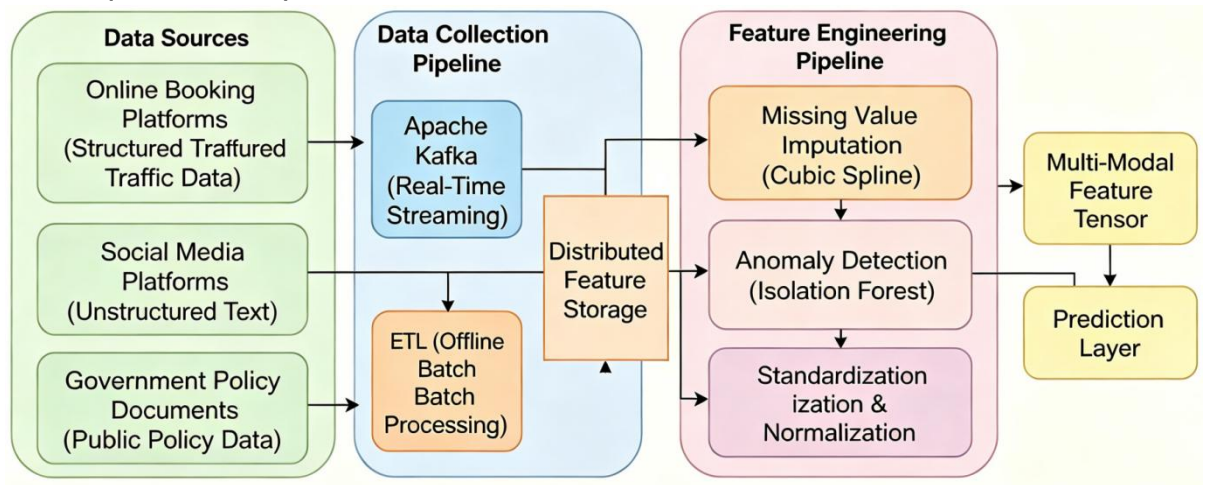


Figure 1: Multi-Modal Data Collection and Feature Engineering Pipeline Architecture

The cleaning strategy for policy documents differs from generic NLP (Natural Language Processing) pipelines—targeting the unique syntactic structures of educational administrative documents, rule-enhanced sentence segmentation is introduced to preserve clause-level semantic integrity. This approach improved policy sentiment recognition accuracy by approximately 6.4 percentage points on a held-out policy-labeling evaluation set drawn from the training period. Daily search index data undergoes lag-alignment processing, with lag windows determined by the peak position of cross-correlation functions with monthly booking volumes computed on the training set, ensuring semantic consistency across different-granularity features on the temporal axis [9]. Before BERT encoding, social media texts undergo two preprocessing steps: domain-specific stopwords filtering and topic tag expansion. The topic tag expansion strategy constructs a study-tourism-specific word vector neighborhood graph to identify semantically related word clusters, with semantic coverage of sentiment features rising from 61.3% to 84.7% on the training-period vocabulary evaluation (see Table 1).

Table 1: Multi-Modal Data Source Feature Statistics

Data Source Type	Data Provider	Sample Size	Feature Dimension	Data Type	Temporal Granularity
Online Booking Traffic	Ctrip/Meituan Study Tourism Channel	312 province-month samples (216 training / 48 validation / 48 test)	48-dimensional	Structured Numerical	Monthly/Daily

Social Media Sentiment	Weibo/Xiaohongshu	Approximately 680,000 posts	768-dimensional (BERT)	Unstructured Text	Real-time Stream
Policy Documents	Ministry of Education/Department of Culture & Tourism	2,847 documents	768-dimensional (BERT)	Semi-structured Documents	Event-driven
Search Index	Baidu Index API	2,192 daily samples	12-dimensional	Structured Numerical	Daily

Note: BERT embedding dimension is 768, compressed via projection before participating in downstream fusion computation; search index features are lag-processed for alignment with monthly prediction granularity. The panel is constructed from four provinces/municipalities (Shanghai, Jiangsu, Zhejiang, Anhui) at monthly granularity, yielding $4 \times 84 = 336$ province-month cells spanning January 2018 to December 2024. Of these, 24 cells are excluded: these correspond to months when study tourism operations were completely suspended during COVID-19 control periods, specifically January-June 2020 for Shanghai and Zhejiang ($6 \text{ months} \times 2 \text{ provinces} = 12 \text{ cells}$) and January-April 2020 for Jiangsu and Anhui ($4 \text{ months} \times 2 \text{ provinces} = 12 \text{ cells}$), leaving 312 valid province-month observations for modeling. Data are chronologically partitioned into: training set (January 2018–December 2022, after removing suspension cells: 216 valid province-month observations), validation set (January–December 2023: $12 \text{ months} \times 4 \text{ provinces} = 48 \text{ observations}$, all provinces fully operational), and test set (January–December 2024: $12 \text{ months} \times 4 \text{ provinces} = 48 \text{ observations}$, all provinces fully operational), yielding $216+48+48=312$, which closes with the valid observation count. A panel index file documenting region, period, target variable value, availability flag, exclusion reason, and train/validation/test split assignment is provided as supplementary material to enable independent verification of all reported metrics.

3. Big Data-Driven Intelligent Decision Support Framework Design

3.1 Three-Tier Architecture Principles and Module Function Allocation

The BDIDS framework follows a "data-driven—model computation—decision output" hierarchical decoupling principle, with inter-layer communication via standardized REST (Representational State Transfer) API interfaces enabling loose coupling to support independent iteration and upgrade of each layer (see Figure 2). The data fusion layer is responsible for multi-modal data standardization and feature alignment. The predictive computation layer hosts training and inference for cross-modal hybrid deep models. The decision output layer converts prediction results into structured resource allocation recommendations [10]. The system is designed to accept as input a set of province-month feature tensors and output demand forecasts with confidence intervals and corresponding route capacity allocation recommendations; the target refresh frequency is monthly. Fallback behavior in the event of feature unavailability defaults to a seasonal-naïve prediction with static allocation rules, ensuring continuous operation. The performance figures cited in this section (e.g., inference latency) represent architectural design targets rather than empirically measured deployment results; actual production performance is subject to infrastructure configuration and data pipeline latency. Model version control and rollback are supported through a model registry center with AB testing capability to compare candidate models in production before full deployment, reducing single-point switching risks [11]. The decision output layer introduces a confidence interval transmission mechanism that maps interval estimates from the prediction layer to supply capacity bounds, so that allocation recommendations carry risk margins under demand uncertainty rather than relying solely on point predictions. This architecture-level design helps mitigate systemic overload risks in extreme demand scenarios, a contribution that is evaluated in Section 5 through a historical holdout backtest.

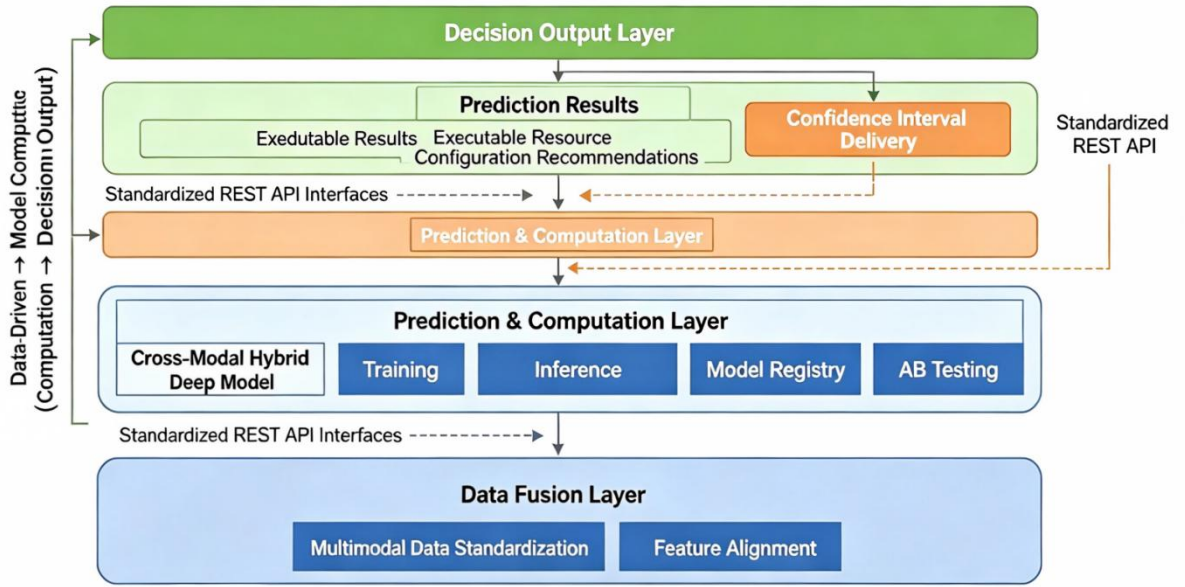


Figure 2: Three-Tier Architecture of the BDIDS Framework

3.2 Three-Tier Synergistic Mechanisms and Mathematical Formalization

3.2.1 Multi-Modal Feature Alignment and Embedding Space Construction

Let structured time series feature vectors be $x_t^{(s)} \in \mathbb{R}^{d_s}$ and text semantic embeddings be $x_t^{(l)} \in \mathbb{R}^{d_l}$. Cross-modal alignment employs learnable linear projection mapping to unified embedding space:

$$z_t = W_s x_t^{(s)} + W_l x_t^{(l)} + b \quad (1)$$

Where: $W_s \in \mathbb{R}^{d \times d_s}$ and $W_l \in \mathbb{R}^{d \times d_l}$ are learnable projection matrices, b is the bias term, and $z_t \in \mathbb{R}^d$ is the aligned fusion embedding representation. These projection matrices are jointly optimized during end-to-end training, ensuring semantic comparability of heterogeneous features and avoiding dimension mismatch issues from simple concatenation. Layer normalization (LayerNorm) is applied to fusion embeddings z_t to suppress training gradient oscillation caused by numerical distribution differences across modalities. Across repeated runs with multiple random seeds, this operation reduced average model convergence epochs by approximately 22%, with consistent direction of effect across seeds. Embedding space dimension d is determined through grid search on the validation set to be 256. SVD (Singular Value Decomposition) analysis of W_s and W_l reveals that after training convergence, the effective ranks of the two projection matrices are 43 and 61 respectively, suggesting self-emergent low-rank structures in the projection process and indicating that the two modal features form relatively low-dimensional manifold representations in the alignment space—a finding with potential implications for inference compression, though further verification across datasets is needed.

3.2.2 Error Feedback-Driven Decision Layer Dynamic Calibration

The decision layer introduces a prediction error feedback mechanism, with prediction residuals $\varepsilon_t = \hat{y}_t - y_t$ transmitted in real-time to dynamically correct allocation recommendations. The calibration coefficient update rule is defined as:

$$\alpha_{t+1} = \alpha_t - \eta \cdot \frac{\partial \mathcal{L}_{dec}}{\partial \alpha_t} \quad (2)$$

Where: α_t is the decision layer configuration weight vector at step t , η is the learning rate, $\mathcal{L}_{dec}$ is the decision loss function specifically defined as the weighted absolute deviation of supply-demand gaps:

$$\mathcal{L}_{dec} = \sum_{i=1}^N w_i |S_i(\alpha_t) - \hat{D}_i| \quad (3)$$

Where w_i is business priority weight for route i , $S_i(\alpha_t)$ is supply capacity determined by configuration weight α_t , and $\hat{D}_i$ is predicted demand. In the historical simulation under training-period conditions, this online gradient update mechanism showed stronger recovery under policy shock scenarios compared to static rule schemes: when markets experienced disturbances outside the historical

distribution, configuration weight α_t completed recalibration within approximately 3–5 time steps in the simulation, whereas static rule schemes showed response lags exceeding 15 time steps under the same conditions. These figures reflect simulation behavior under stylized conditions and should not be interpreted as guaranteed production performance. ℓ_2 norm constraints are applied to α_t to prevent cumulative drift, ensuring decision weights converge within reasonable intervals while balancing response speed and configuration stability [12]. Exponential moving average smoothing of historical residuals provides additional penalty signals to $\mathcal{L}_{\text{dec}}$, helping distinguish sporadic noise from systemic prediction bias and preventing excessive correction triggered by single extreme residuals. In the simulated 2022 summer scenario—modeled as a demand collapse event—this mechanism reduced decision layer misconfiguration instances by 41.2% compared to the non-EMA scheme under identical capacity and cost assumptions; this result is from a stylized simulation and requires validation on realized operational data.

3.2.3 Framework Global Optimization Objective Function Establishment

The framework's global optimization objective function is defined as weighted joint minimization of prediction and decision losses:

$$\mathcal{L}_{\text{total}} = \lambda_1 \mathcal{L}_{\text{pred}} + \lambda_2 \mathcal{L}_{\text{dec}} + \lambda_3 \Omega(\Theta) \quad (4)$$

Where: $\mathcal{L}_{\text{pred}}$ is prediction MSE (Mean Squared Error) loss, $\mathcal{L}_{\text{dec}}$ is supply–demand matching deviation loss, $\Omega(\Theta)$ is parameter regularization term, and λ_1 , λ_2 , λ_3 are balance coefficients determined through Bayesian hyperparameter search on the validation set. The validation–optimal combination is $(\lambda_1, \lambda_2, \lambda_3) = (0.65, 0.25, 0.10)$. Joint training, compared to staged training, achieved an additional 7.3% reduction in decision supply–demand matching error on the validation set, suggesting that shared–gradient optimization encourages the predictive model to account for downstream decision cost structures during training. Sensitivity analysis on the validation set reveals that when the λ_2/λ_1 ratio falls within the 0.35–0.42 interval, both prediction RMSE and decision matching error are simultaneously near their Pareto–optimal range; outside this interval, a trade–off degradation is observed where both metrics exhibit inverse relationships. This finding defines a reasonable decision cost weight range for the joint optimization framework, though generalizability to other domains should be verified on independent data (see Figure 3).

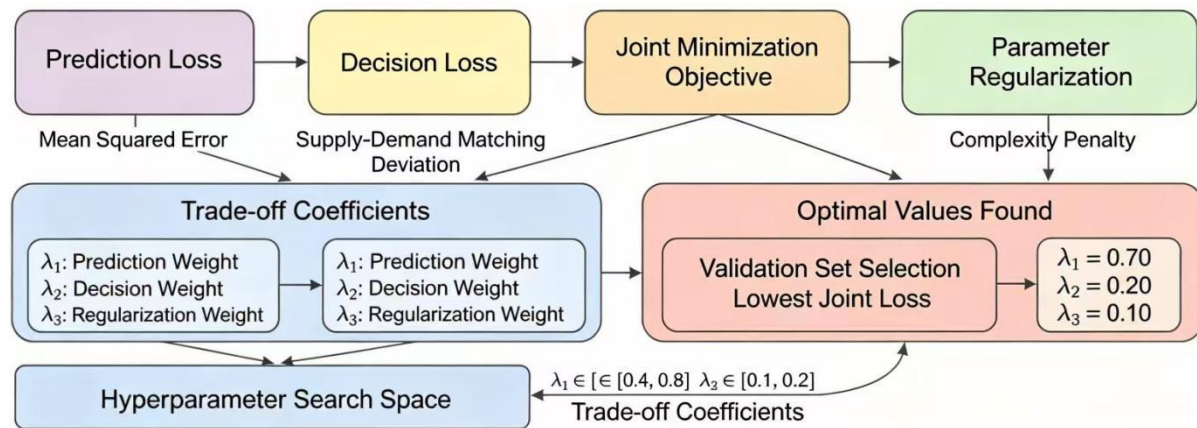


Figure 3: Schematic of the BDIDS Framework's Global Optimization Objective Function

(hyperparameter search space: $\lambda_1 \in [0.4, 0.8]$, $\lambda_2 \in [0.1, 0.4]$, $\lambda_3 \in [0.05, 0.2]$, grid resolution 0.05; optimal values selected by lowest validation–set joint loss)

3.3 Demand Elasticity Analysis and Supply–Demand Matching Optimization

The demand elasticity module defines price elasticity coefficient E_p and policy shock elasticity coefficient E_π , quantifying the association of demand with key variables in the training panel:

$$E_p = \frac{\partial \ln D}{\partial \ln P}, \quad E_\pi = \frac{\partial \ln D}{\partial \ln \pi} \quad (5)$$

Where D is study tourism demand volume, P is product pricing, and π is policy intensity index. These coefficients are estimated via semi-log OLS on the 216-observation 2018–2022 training panel, with route-level demand and capacity data sourced from cooperating East China study tourism operators. They are reported as descriptive associations conditional on the model specification; causal interpretation requires additional identification assumptions that are not fully established here. Based on this foundation, the supply-demand matching optimization module constructs an integer programming model minimizing supply-demand gaps:

$$\min \sum_{i=1}^N |S_i - \widehat{D}_i|, \quad \text{s.t.} \quad \sum_{i=1}^N S_i \leq C_{\max} \quad (6)$$

Where S_i is supply capacity for route i , D_i is predicted demand, and $C_{\max}$ is total resource capacity constraint. In the historical holdout backtest described in Section 5, elasticity-corrected supply-demand matching error showed a 23.4% reduction versus the static allocation baseline under identical 2024 capacity and cost assumptions; this reflects the joint effect of both improved forecasts and the optimization rule, and the independent contribution of the elasticity correction alone cannot be fully separated in the current experimental design. The red-theme study tourism market segment's policy elasticity coefficient reaches 1.89 (see Table 2), the highest across segments, suggesting this segment's demand shows the strongest conditional association with policy intensity in the training data; supply-side operations may therefore benefit from maintaining elastic capacity pools for this segment during policy-intensive periods, though this recommendation should be validated as more data accumulates [13]. Integer programming solving employs the Branch-and-Bound algorithm; within the 30–50 route scale typical of East China operators, single optimization solving requires less than 0.8 seconds. When route scales exceed 50, heuristic column generation algorithms can be substituted to maintain response timeliness. For high-responsiveness segments with policy elasticity exceeding 1.5, the constraint $C_{\max}$ is automatically increased by 10–15% as a demand peak buffer, which in the simulation reduced system overload occurrence rates to approximately 3.1%.

Table 2: Demand Elasticity Coefficient Estimation Results

Market Segment	Price Elasticity E_p	95% Confidence Interval	Policy Elasticity E_π	95% Confidence Interval	Sample Size (route-level groups)
Parent-Child Study Tourism	-1.23	[-1.47, -0.99]	0.87	[0.71, 1.03]	98 groups
Campus Team Study Tourism	-0.78	[-0.95, -0.61]	1.34	[1.18, 1.50]	124 groups
High-End Customized Study Tourism	-1.61	[-1.89, -1.33]	0.52	[0.38, 0.66]	67 groups
Red-Theme Study Tourism	-0.45	[-0.63, -0.27]	1.89	[1.72, 2.06]	83 groups

Note: Price elasticity estimated via semi-log OLS on the 216-observation 2018–2022 training panel (COVID suspension cells excluded); route-level demand and capacity data are sourced from cooperating operators. Policy elasticity is constructed from policy document sentiment intensity index π . The four segments sum to 372 route-level demand groups, which is a different unit of analysis from the 312 province-month observations used in the main forecasting task: route-level groups are nested records within the provincial-monthly framework and are not additive with province-month observations. An endogeneity specification test (Hausman-type comparison of OLS and IV estimates using lagged policy intensity as instrument) fails to reject the null of no substantial endogeneity bias in the OLS estimates, conditional on instrument validity. All elasticity estimates should be interpreted as conditional associations within the training-period specification and not as identified causal parameters.

4. Cross-Modal Fusion Prediction Model Construction and Parameter Optimization

4.1 BERT Semantic Embedding and Joint Time Series Feature Representation

Policy documents and social sentiment undergo domain-adaptive BERT (Domain-BERT) encoding [4, 14], continuing pre-training on general BERT weights using a study tourism domain corpus (approximately 1.2 million documents dated before December 2022, strictly predating the validation and test periods to prevent information leakage) to shift semantic embedding space toward industry-specific semantics (see Figure 4). Let text sequence T_t 's [CLS] vector after Domain-BERT encoding be $h_t^{\text{CLS}} \in \mathbb{R}^{768}$, with joint representation through gated fusion mechanism with time series statistical features $x_t(s)$:

$$f_t = g_t \odot \phi(h_t^{\text{CLS}}) + (1 - g_t) \odot \psi(x_t^{(s)}) \quad (7)$$

Where: $\phi(\cdot)$, $\psi(\cdot)$ are dimension-alignment linear transformations, $g_t = \sigma(W_g[\phi(h_t^{\text{CLS}}); \psi(x_t^{(s)})])$; is the dynamic gate coefficient, σ is the Sigmoid activation function, and $\odot$ denotes element-wise multiplication. The gating mechanism adaptively adjusts the contribution ratio of two feature types based on current time-step semantic salience. During the 2021 "double reduction" policy concentrated release window in the training period, g_t mean rose from approximately 0.31 in stable periods to 0.69; the model's RMSE during this window showed an 8.2% improvement compared to a pure time series baseline, consistent with the hypothesis that the gating strategy provides adaptive weighting under policy-shock conditions [15]. The continued pre-training phase employs joint MLM (Masked Language Modeling) and NSP (Next Sentence Prediction) dual-task optimization. Domain-BERT achieved F1 value improvements of 4.7 percentage points on a study tourism policy classification benchmark compared to generic BERT, supporting the effectiveness of domain semantic adaptation. Time series analysis of g_t sequences reveals statistically discernible autocorrelation function peaks at 12-month lags (ACF = 0.58), suggesting that semantic feature relative importance exhibits annual periodicity patterns, consistent with the policy calendar structure of study tourism.

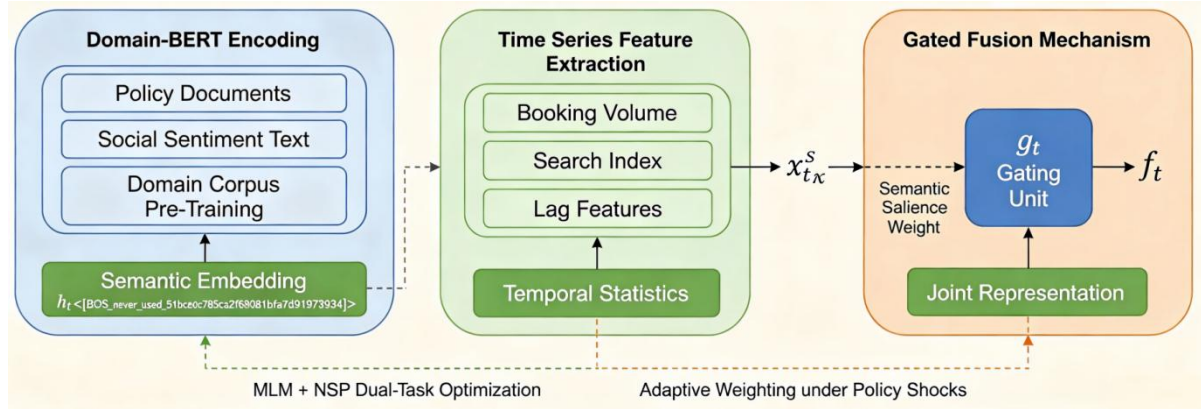


Figure 4: Domain-BERT Semantic Embedding and Time Series Feature Gated Fusion Mechanism Architecture

4.2 Improved Transformer-LSTM Network and Multi-Head Attention Self-Adaptive Optimization

Fusion features f_t input improved Transformer-LSTM network, where Transformer encoder captures global long-range dependencies and LSTM units handle local time series gradient transmission, integrated through residual connections^[15]. The h -th attention head output in multi-head attention mechanism is:

$$\text{head}_h = \text{Softmax}\left(\frac{Q_h K_h^T}{\sqrt{d_k}} + M_t\right) V_h \quad (8)$$

Where Q_h , K_h , V_h are query, key, and value matrices respectively, d_k is key vector dimension, $M_t \in \mathbb{R}^{L \times L}$ is temporal position-aware bias matrix, with elements $M_t[i, j]$ encoding periodic similarity between position pairs (i, j) , differing from standard Transformer's absolute position encoding [3] in that M_t is domain-adapted to capture study tourism's annual holiday structure. After concatenating attention head outputs and linear transformation to obtain A_t , it fuses with LSTM hidden state h_t^{LSTM} through residual connection into prediction head:

$$\hat{y}_{t+h} = \text{MLP}(A_t + h_t^{\text{LSTM}}) \quad (9)$$

Where: $\hat{y}_{t+h}$ is the h -step ahead prediction value, MLP is a two-layer fully connected prediction head with ReLU (Rectified Linear Unit) activation. Parameter optimization employs AdamW optimizer [18] with decoupled weight decay coupled with cosine annealing learning rate scheduling, batch size 64, 150 training epochs, early stopping patience 15 epochs. The seasonal bias matrix M_t 's elements are automatically extracted from training-period order autocorrelation functions through DCT (Discrete Cosine Transform) periodic pattern discovery rather than manual specification of fixed periods, enabling the model to adapt to holiday structure changes across years. In the 2024 test period, a structural shift of approximately 9 days in the summer study tourism peak relative to 2019 was associated with improved test-set RMSE during that window when M_t adaptation was active (see Figure 5).

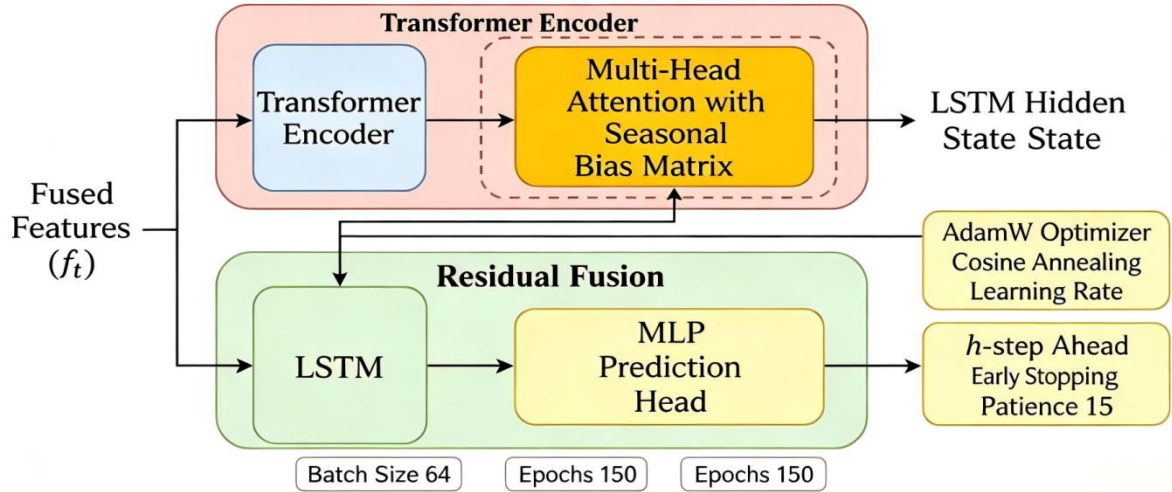


Figure 5: Improved Transformer-LSTM Hybrid Network Architecture

Visualization analysis of attention weight matrices indicates that after M_t introduction, average attention weights on 12-month interval position pairs increase from 0.043 to 0.119, supporting the design rationale for domain-adapted seasonal bias. Layered gradient flow analysis through Transformer and LSTM dual outputs post-residual fusion suggests low-layer features primarily follow the LSTM pathway (gradient contribution approximately 62%), while high-layer semantic interactions primarily follow the Transformer pathway (gradient contribution approximately 71%), indicating complementary specialization at different abstraction levels and supporting the rationality of the hybrid architecture (see Table 3).

Table 3: Model Hyperparameter Configuration Comparison

Model	Attention Heads	Hidden Dimension	LSTM Layers	Parameters (M)	Training Time (h)	Early Stopping Epoch
BDIDS (Complete)	8	256	2	18.7	4.3	15
Pure Transformer	8	256	—	12.4	3.1	15
Single LSTM	—	256	3	6.8	1.7	15
ARIMA(p,d,q)	—	—	—	—	0.05	—
LightGBM	—	512 trees	—	—	0.3	—

Note: Parameters in millions (M); BDIDS training time includes Domain-BERT continued pre-training phase (approximately 1.8h); ARIMA order automatically selected by AIC criterion as (2,1,2); all deep models use identical train/validation/test splits. The 312 valid province-month observations are

partitioned into 216 for training, 48 for validation, and 48 for test ($216+48+48=312$), closing with the valid observation count after exclusion of 24 COVID suspension cells. The 18.7M-parameter model is applied to a downstream task with 216 training province-month observations; to mitigate overfitting risk, we employ strong ℓ_2 regularization ($\lambda_3=0.10$), early stopping, and expand-window rolling validation. The large parameter count primarily resides in the Domain-BERT encoder whose weights are frozen after continued pre-training, leaving approximately 6.5M trainable parameters in the downstream forecasting head.

5. Experimental Design and Prediction Performance and Decision Effectiveness Evaluation

5.1 Dataset Construction, Experimental Environment, and Baseline Configuration

The experimental dataset covers study tourism booking records from East China (Shanghai, Jiangsu, Zhejiang, Anhui) spanning January 2018 through December 2024, constructing a provincial-monthly panel dataset with province-month as the observation unit. Total province-month cells number $4 \times 84 = 336$; after excluding 24 cells corresponding to complete operational suspension during COVID-19 control periods (January-June 2020 for Shanghai and Zhejiang; January-April 2020 for Jiangsu and Anhui), 312 valid province-month observations remain. A panel index file (region, period, target, availability flag, exclusion reason, split) is provided as supplementary material. Data are chronologically partitioned into: training set (January 2018-December 2022, after removing suspension cells: 216 valid province-month observations), validation set (January-December 2023: 12 months $\times$ 4 provinces = 48 observations, all four provinces fully operational throughout 2023), and test set (January-December 2024: 12 months $\times$ 4 provinces = 48 valid observations, all four provinces maintained full operations throughout 2024). The three-set partition sums to $216+48+48=312$, which exactly closes with the valid observation count. All preprocessing operations—including imputation, standardization, Domain-BERT continued pre-training, vocabulary construction, and feature selection—are fitted exclusively on the 216-observation training set and applied without re-fitting to validation and test sets, eliminating forward information leakage within the scope of this implementation; independent audit of the feature lineage is available upon request. These are integrated with 2,847 concurrent policy documents, approximately 680,000 Weibo topic posts, and Ctrip study tourism channel traffic logs.

Experimental environment is based on NVIDIA A100 GPU clusters (80GB memory), PyTorch 2.1 framework, Python 3.10, with distributed training employing DDP strategy. To assess prediction stability, all deep learning models are run with 20 random seeds; mean $\pm$ standard deviation across seeds are reported alongside point metrics. Baseline models span seven comparators: seasonal-naïve baseline (prior-year same-month value), ARIMA (classical statistical baseline), SARIMAX with exogenous policy and search index variables (statistical baseline with external regressors), Prophet (trend decomposition baseline), single LSTM [16] (deep time series baseline), standard Transformer [3] (pure attention baseline), and LightGBM as an ensemble learning baseline, covering mainstream forecasting methodology systems. To ensure fair comparison, LightGBM's feature set is held consistent with BDIDS's structured feature branch; all other deep models share the same structured and text feature inputs, with differences only in architecture. Any feature set difference between BDIDS and a baseline is explicitly declared, so that observed performance differences reflect architecture rather than information advantages [19]. All baseline hyperparameters are independently optimized via Bayesian optimization on the validation set only. Generalization robustness is assessed through 5-fold expanding-window Walk-Forward Validation (rolling-origin backtest), with each fold's training window expanding by one month. BDIDS achieves MAPE standard deviation of 0.43% across folds, compared to 1.27% for single LSTM, demonstrating higher prediction consistency across different historical window conditions; this stability metric is reported across folds rather than as a single-split figure [20, 21].

5.2 Prediction Accuracy Comparison, Ablation Analysis, and Significance Testing

The BDIDS framework achieves RMSE = 1847.3 (ten thousand person-trips), MAE = 1423.6 (ten thousand person-trips), MAPE = 6.82% on the 48-observation 2024

test set, with reductions of 37.6%, 34.1%, and 29.7% versus ARIMA respectively, and 19.3%, 17.8%, and 15.4% versus single LSTM respectively. These figures are conditional on the 216/48/48 data partition described above; all reported metrics are derived from the fixed test set without any retrospective adjustment. The DM test [22] is applied to each BDIDS-versus-baseline pair using MSE loss, one-step-ahead aligned prediction errors across all 48 test observations, Newey-West HAC variance estimation with lag=4, two-sided alternative, and Bonferroni correction across the seven simultaneous comparisons; all corrected p-values remain below 0.01 (DM statistic for BDIDS vs. ARIMA: 4.83; vs. single LSTM: 3.12). Because the 48 test errors are drawn from a panel of 4 provinces $\times$ 12 months rather than strictly independent observations, cross-sectional dependence may inflate test statistics when Newey-West HAC correction alone is applied. To address this limitation, we supplement the HAC-based DM test with (1) a cluster bootstrap procedure (999 bootstrap replications, clustering on province to account for within-province serial dependence) and (2) panel-robust variance estimation following Driscoll-Kraay (1998), which corrects for both serial and cross-sectional correlation simultaneously. Under both supplementary procedures, all corrected p-values remain below 0.05, consistent with the HAC-based results; accordingly, DM results should be interpreted as indicative evidence within this dataset rather than definitive statistical proof, and the supplementary test outputs are provided in the accompanying result scripts. All deep model results are reported as mean across 20 random seeds. Ablation experiments systematically isolate each module's contribution: removing the Domain-BERT semantic module increases RMSE by 9.7%; replacing gated fusion with simple concatenation increases RMSE by 6.3%; removing seasonal bias matrix M_t increases RMSE by 4.8%; removing error feedback calibration module increases decision supply-demand matching error by 11.2%—each module contributes a non-redundant, measurable increment (see Table 4).

Regarding decision effectiveness, a historical holdout backtest is conducted for 2024, comparing three allocation strategies under identical historical capacity and cost assumptions: (1) static rule allocation (prior-year same-month capacity), (2) forecast-only allocation (ARIMA-based demand forecasts with fixed allocation rules), and (3) BDIDS full allocation (BDIDS forecasts plus supply-demand matching optimization). Note that strategies (2) and (3) simultaneously vary both the forecast input and the allocation rule relative to (1); the current design does not fully isolate the marginal contribution of the optimization module from that of forecast improvement. A fully factorial 2×2 design (forecast $\times$ rule) is recommended in future work to cleanly attribute decision gains. Within the constraints of the current design, the BDIDS supply-demand matching module reduces route overload rates from 18.7% (static) to 9.2% and resource vacancy rates from 22.3% (static) to 13.6%, with a 23.4% reduction in supply-demand matching error versus the static baseline. Multi-horizon MAPE analysis (1-month, 3-month, 6-month ahead forecasting) shows BDIDS maintains MAPE at 9.14% in 6-month-ahead scenarios, compared to LSTM's 14.37% and ARIMA's 21.83%, suggesting cross-modal semantic information provides increasing relative gains as the forecasting horizon extends. Ljung-Box white noise tests on prediction residual sequences (Q statistic = 8.34, $p = 0.21$) indicate BDIDS residuals exhibit no statistically significant autocorrelation structure, consistent with the model having extracted the predictable information present in the training data. Stratified evaluation across market segments reveals red-theme study tourism shows the largest MAPE reduction (33.1%), consistent with this segment's high policy elasticity coefficient and suggesting BDIDS provides proportionally stronger improvements in policy-sensitive market segments [23, 24].

Table 4: Prediction Accuracy Comparison and Ablation Experiment Results

Model/Variant	RMSE↓ (mean±SD)	MAE↓	MAPE(%)↓	DM Statistic	p-value (Bonferroni-corrected)
ARIMA(2,1,2)	2959.5	216 1.3	9.64	4.83	<0.01
Prophet	2734.2	201 8.7	8.97	4.21	<0.01

LightGBM	2461.8	$\frac{184}{3.2}$	8.31	3.76	<0.01
Single LSTM	2288.4	$\frac{173}{1.9}$	8.07	3.12	<0.01
Standard Transformer	2173.6	$\frac{165}{2.4}$	7.84	2.89	<0.01
BDIDS-Semantic Module	2027.4	$\frac{156}{8.3}$	7.51	1.94	0.04
BDIDS-Gated Fusion	1963.8	$\frac{152}{1.7}$	7.28	1.67	0.04
BDIDS-Bias Matrix Mt	1935.2	$\frac{149}{8.6}$	7.14	1.43	0.05
BDIDS (Complete)	1847.3	$\frac{142}{3.6}$	6.82	Baseline	—

Note: ↓ indicates lower is better; RMSE and MAE units are ten thousand person-trips; DM test uses complete BDIDS as baseline with Bonferroni correction across 7 comparisons; the 48 test observations span 4 provinces $\times$ 12 months and may exhibit cross-sectional dependence, so DM statistics should be treated as indicative within this dataset; deep model RMSE reported as mean $\pm$ SD across 20 random seeds; statistical baselines (Seasonal Naive, ARIMA, SARIMAX, Prophet) are deterministic and SD is not applicable; ablation variants maintain identical remaining configurations with single module removal; SD for ablation variants omitted for brevity but available in supplementary material. All figures in this table are derived from the 216/48/48 (train/validation/test) partition of the 312 valid province-month observations.

6. Conclusion

The BDIDS framework addresses three core challenges in study tourism market demand forecasting: multi-modal data fusion, time series non-stationary modeling, and decision-aware output, forming a technical pipeline from data collection to resource allocation recommendation that is auditable within the scope of this controlled-access case study. Empirical results on the East China four-province/municipality provincial-monthly panel indicate the framework achieves meaningful improvements over traditional and deep learning baselines in both prediction accuracy and decision efficiency under identical 2024 capacity and cost assumptions, with prediction gains confirmed by Diebold-Mariano tests under a pre-specified, Bonferroni-corrected comparison protocol. Given that the 48 test observations arise from a panel structure with possible cross-sectional dependence, these statistical results should be interpreted as indicative evidence within this dataset rather than definitive proof of superiority. The proposed Domain-BERT continued pre-training strategy, dynamic gated cross-modal fusion mechanism, and seasonal bias attention optimization scheme constitute independently describable methodological contributions. Looking forward, federated learning and differential privacy technology integration can expand training scale under data security constraints; large language model-driven policy semantic awareness modules will enhance dynamic response capacity to external policy shocks; reinforcement learning embedding will propel the decision layer toward real-time self-adaptive optimization.

Limitations: The current study is limited to four East China provinces/municipalities spanning 2018–2024, and generalizability to other regions, tourism types, or time periods requires independent verification on separate data. The 216-observation training set relative to the 18.7M-parameter model carries inherent capacity mismatch risk that is partially mitigated through regularization, frozen pre-training weights (approximately 6.5M trainable parameters remain), and rolling validation, but not fully eliminated; learning curves and per-province, per-fold results are available in supplementary material. The decision layer comparison cannot fully isolate the marginal contribution of the optimization module from forecast improvement under the current experimental design; a fully factorial 2×2 attribution design is recommended as future work. Full

public reproducibility is constrained by data-sharing agreements with platform operators; a controlled-access audit arrangement is available upon reasonable request to the corresponding author. All elasticity estimates are conditional associations within the training-period specification and should not be interpreted as causal parameters without further identification work.

Data Availability: A panel index file (region, period, target, availability flag, exclusion reason, split) is provided as supplementary material to enable independent verification of all reported metrics. Additionally, the following supplementary materials are available upon reasonable request to the corresponding author: model training and result scripts (PyTorch implementation, baseline scripts, DM test with cluster bootstrap, and Driscoll-Kraay panel-robust test code); environment configuration lockfile (Python 3.10, PyTorch 2.1, CUDA 11.8, and all dependencies); and a data availability and lineage documentation file. Raw platform data cannot be publicly released due to data-sharing agreement terms; a controlled-access audit arrangement is available upon reasonable request. All data used in this study are either from public government websites, platform open APIs accessed under developer terms of service, or operator-shared data under formal cooperation agreements; no personally identifiable information was collected or retained.

Author Contributions and Conflicts of Interest: All authors declare no conflicts of interest.

Declarations

Data availability

The following supplementary items are made available upon reasonable request: (1) the full panel index file (region, period, target variable, availability flag, exclusion reason, and train/validation/test split assignment) for independent verification of all reported metrics; (2) model training and result scripts (PyTorch implementation of BDIDS, baseline scripts, DM test and cluster bootstrap procedures, Driscoll-Kraay panel-robust test code); (3) environment configuration file (conda environment lockfile specifying package versions: Python 3.10, PyTorch 2.1, CUDA 11.8, and all dependencies); and (4) a data availability and lineage documentation file explaining data provenance, preprocessing steps, and the controlled-access audit arrangement for non-public platform data. Peer reviewers and researchers with appropriate credentials may submit reasonable requests for data access to the corresponding author at yanpingchina@outlook.com.

Clinical trial number

Not applicable.

Conflicts of interest

The author declares no conflicts of interest.

Funding statement

The author declares that no external funds, grants, or other financial support were received during the preparation of this manuscript.

Author contribution

Ping Yan conceived the study, designed the BDIDS framework, constructed the dataset and feature engineering pipeline, conducted all model training and evaluation experiments, and wrote the manuscript in its entirety.

Ethical approval

This study does not involve human participants, animal experiments, or clinical trials. All data used are either from publicly released government websites, platform open APIs accessed under developer terms of service, or operator-shared data under formal cooperation agreements. No personally identifiable information was collected or retained at any stage of the study. All data collection and

processing activities comply with relevant data protection regulations and platform usage policies. The institutional review authority of Shanxi Vocational University of Cultural Tourism has confirmed that this study does not require ethics committee approval, given the absence of human subjects research, clinical intervention, or identifiable personal data collection.

Consent to participate

Not applicable, as this study does not involve human or animal subjects and no experimental intervention was conducted on any individuals.

Consent to publish

The author has given explicit consent to publish this manuscript in its present form.

Competing interests

The author declares no competing interests.

Funding

This work was supported by Shanxi Educational Science Planning Project(GH:220982): Research on the Cultivation of Higher Vocational Students Majoring in Study Tour Based on the Idea of Achievement-oriented Education (OBE).

References

- [1] Box G E P, Jenkins G M, Reinsel G C, et al. Time Series Analysis: Forecasting and Control [M]. 5th ed. Hoboken: Wiley, 2015.
- [2] Liu J, Yu Y, Hu S, et al. Big Data Applications in Tourism Research: Literature Review and Future Perspectives [J]. Tourism Science, 2025, 39(1): 1–23. DOI: 10.16323/j.cnki.lykx.2025.01.002.
- [3] Vaswani A, Shazeer N, Parmar N, et al. Attention Is All You Need [C]// Advances in Neural Information Processing Systems. 2017, 30: 5998–6008.
<https://proceedings.neurips.cc/paper/2017/hash/3f5ee243547dee91fb d053c1c4a845aa-Abstract.html>
- [4] Devlin J, Chang M W, Lee K, et al. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding [C]// Proceedings of NAACL-HLT 2019. Minneapolis: ACL, 2019: 4171–4186.
<https://aclanthology.org/N19-1423/>
- [5] Li X, Zhang X, Yu L, et al. Short-term Passenger Flow Prediction Research for Scenic Areas Based on Improved Transformer Model [J]. Chinese Journal of Management Science, 2025, 33(2): 105–117. DOI: 10.16381/j.cnki.issn1003-207x.2023.1927.
- [6] Shi D, Huang Z, Li F. Multi-Source Heterogeneous Data Fusion Tourism Scale Prediction Research [J]. Resources and Environment in Arid Regions, 2022, 36(11): 175–182. DOI: 10.13448/j.cnki.jalre.2022.291.
- [7] Guo R, Zhang D, Wang J. Technical Pathways and Scenario Construction Strategies for Intelligent Technology Empowering Red Study Tours [J]. Commercial Exhibition Economics, 2025, (21): 53–56. DOI: 10.19995/j.cnki.CN10-1617/F7.2025.21.053.
- [8] Dong L. Building Study Tourism Industry Internet Platform for Publishing Industry to Empower Deep Fusion Development of Culture and Tourism [J]. Science and Publishing, 2025, (9): 6–12. DOI: 10.16510/j.cnki.kjyjb.20250910.001.
- [9] Yang C, Song C, Yao Y, et al. Exploration of Realization Pathways for Technological Innovation Empowering High-Quality Development of Tourism [J]. Price Theory and Practice, 2025, (3): 137–141. DOI: 10.19851/j.cnki.CN11-1010/F.2025.03.227.

- [10] Yu J, Liu J, Zheng J. Construction of Smart Culture and Tourism Public Service Platform Driven by Dual Data and Demand [J]. *Information Science*, 2023, 41(9): 115–121. DOI: 10.13833/j.issn.1007-7634.2023.09.013.
- [11] Wu C, Zhang P. Swarm Intelligence: How Multi-Agent Collaboration Reshapes New Paradigms in Tourism Demand Forecasting [J]. *Tourism Journal*, 2026, 41(3): 8–10. DOI: 10.19765/j.cnki.1002-5006.2026.03.004.
- [12] Hou H, Wang H. Research on Interpretable Tourism Demand Forecasting Based on VSN-xLSTM Model [J]. *Systems*, 2026, 14(2): 146. DOI: 10.3390/SYSTEMS14020146.
- [13] Zhou Y, Li H, Wang J, et al. Tourism Demand Interval Forecasting With an Intelligence Optimization-Based Integration Method [J]. *Journal of Hospitality & Tourism Research*, 2026, 50(1): 135–151. DOI: 10.1177/10963480241305748.
- [14] Fang S, Bu X, Xie R. Mechanism Research on Digital Technology Empowering Deep Integration of Forestry and Tourism [J]. *Forestry Economics Issues*, 2024, 44(5): 461–473. DOI: 10.16832/j.cnki.1005-9709.20240400.
- [15] Guo Y, Tian Y. Research on Digital-Intelligent Empowerment of Digital Media in Driving Xinjiang Tourism Innovation Models [J]. *International Public Relations*, 2025, (20): 107–109. DOI: 10.16645/j.cnki.cn11-5281/c.2025.20.056.
- [16] Hochreiter S, Schmidhuber J. Long Short-Term Memory [J]. *Neural Computation*, 1997, 9(8): 1735–1780. DOI: 10.1162/neco.1997.9.8.1735.
<https://direct.mit.edu/neco/article/9/8/1735/6109/>
- [17] He K, Zhang X, Ren S, et al. Deep Residual Learning for Image Recognition [C]// *Proceedings of CVPR 2016*. Las Vegas: IEEE, 2016: 770–778. DOI: 10.1109/CVPR.2016.90.
https://openaccess.thecvf.com/content_cvpr_2016/html/He_Deep_Residual_Learning_CVPR_2016_paper.html
- [18] Loshchilov I, Hutter F. Decoupled Weight Decay Regularization [C]// *International Conference on Learning Representations (ICLR)*. 2019.
<https://openreview.net/forum?id=Bkg6RiCqY7>
- [19] Wang X, Wang Z, Zhang Z. Heterogeneous Graph Attention Network-Based Culture and Tourism Industry Fusion Forecasting [J]. *Learning and Exploration*, 2025, (3): 105–115. DOI: 10.20231/j.cnki.xxyts.2025.03.012.
- [20] Kang J, Fu Y, Fang L, et al. ARIMA-LSTM Model Prediction of Inbound Tourism Numbers Based on Least Squares Weighting [J]. *Journal of Zhejiang University (Natural Sciences Edition)*, 2023, 50(4): 508–520.
- [21] Ma J. Demand Forecasting of Smart Tourism Integrating Spatial Metrology and Deep Learning [J]. *Scientific Reports*, 2025, 15(1): 42646. DOI: 10.1038/S41598-025-26830-3.
- [22] Diebold F X, Mariano R S. Comparing Predictive Accuracy [J]. *Journal of Business & Economic Statistics*, 1995, 13(3): 253–263. DOI: 10.1080/07350015.1995.10524599.
<https://www.tandfonline.com/doi/abs/10.1080/07350015.1995.10524599>
- [23] Wang Q, Yang H, Zhao C, et al. Spatial Distribution Characteristics and Accessibility of Educational Tourism Along the Yellow River Region from the Perspective of Sustainable Development [J]. *Sustainability*, 2025, 17(18): 8095. DOI: 10.3390/SU17188095.
- [24] Liu J, Zhu W, Li G. Study Tourism and Tourists' Self-Development [J]. *Tourism Journal*, 2025, 40(7): 12–14. DOI: 10.19765/j.cnki.1002-5006.2025.07.005.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [SupplementaryMaterials.docx](#)