

International Journal of Reasoning-based Intelligent Systems

ISSN online: 1755-0564 - ISSN print: 1755-0556

<https://www.inderscience.com/ijris>

A general-purpose tourism intent understanding and route planning based on multimodal spatio-temporal pre-training

Qiong Zhang

DOI: [10.1504/IJRIIS.2026.10076976](https://doi.org/10.1504/IJRIIS.2026.10076976)

Article History:

Received:	03 January 2026
Last revised:	05 February 2026
Accepted:	06 February 2026
Published online:	26 March 2026

A general-purpose tourism intent understanding and route planning based on multimodal spatio-temporal pre-training

Qiong Zhang

College of Business Administration,
Zhengzhou University of Science and Technology,
Zhengzhou 450000, China
Email: zhangqiong2023@sina.com

Abstract: Accurately understanding tourists' travel intentions and planning routes has become a key issue in smart tourism. Traditional methods rely on a single data modality and are unable to cope with multi-dimensional user demands in complex temporal and spatial contexts. Therefore, this paper first explicitly models the multi-modal spatio-temporal relationships by constructing timestamp-aligned heterogeneous graphs, and uses graph convolution networks for semantic interaction. A dynamic weight mechanism based on the Pearson correlation coefficient is introduced to optimise the fusion process of cross-modal features, ultimately achieving the precise identification of travel intentions. Finally, guided by intent understanding results, a multi-objective tourism route planning method is designed. Combining heuristic search and evolutionary strategies, an optimal and balanced route plan is generated. Experimental results show that the recognition accuracy of the proposed method reaches 93.51%, and the reverse generation distance of the route planning is only 1.9, demonstrating strong generalisation ability and practical value.

Keywords: travel intent recognition; route planning; pre-trained model; spatio-temporal feature extraction; multimodal feature fusion.

Reference to this paper should be made as follows: Zhang, Q. (2026) 'A general-purpose tourism intent understanding and route planning based on multimodal spatio-temporal pre-training', *Int. J. Reasoning-based Intelligent Systems*, Vol. 18, No. 10, pp.88–106.

Biographical notes: Qiong Zhang is an Associate Professor in the School of Business Administration at Zhengzhou University of Science and Technology, China. She obtained a Master's degree from Yanbian University (2011). She published an EI index paper and two Chinese cores papers. Her research interests include the tourism sustainable development and multimodal feature fusion.

1 Introduction

As the digital economy and cultural tourism industry deeply integrating, personalisation and intelligence have become core directions for upgrading tourism services. Accurately capturing user tourism intent and generating route plans that meet demand are key factors in improving service quality and enhancing user experience (Wu et al., 2017). However, in current tourism service scenarios, users' needs show characteristics such as multimodality, fragmentation, and strong spatiotemporal correlations (Liu et al., 2024). Users not only express their demands through textual descriptions but also frequently provide multimodal information such as image references and voice commands. At the same time, these demands are influenced by temporal constraints and spatial distribution factors, presenting significant challenges to traditional methods of tourism intent understanding and route planning. Traditional tourism intent understanding approaches often rely on single-modal text-based keyword matching or shallow semantic analysis, which struggle to effectively integrate hidden needs from multimodal

information such as images and voice (Huang et al., 2023). Moreover, these methods inadequately exploit spatiotemporal correlation features in user demands, leading to low accuracy in intent recognition and poor generalisation capabilities (Yan et al., 2023). At the route planning level, existing approaches are mostly based on pre-defined attraction weights or fixed optimisation goals. They lack dynamic adaptation to personalised user intents and struggle to efficiently handle large-scale spatiotemporal data, resulting in suboptimal practicality and timeliness of planning outcomes.

Recently, the rapid development of multimodal learning and pre-training technologies has provided new insights for addressing the above-mentioned issues (Miyazawa et al., 2022). Multimodal pre-trained models can effectively uncover common features from multi-source information such as text, images, and voice through cross-modal semantic alignment, achieving unified representation and understanding of multimodal data (Wang et al., 2023). Meanwhile, spatiotemporal pre-training technologies

capture temporal dependencies and spatial correlation patterns by pre-learning large-scale spatiotemporal data, enhancing the model's adaptability to dynamic spatiotemporal changes (Arif and Bashir, 2024). Combining multimodal learning with spatiotemporal pre-training technologies to construct a generalisable tourism intent understanding and route planning model has the potential to overcome the modal limitations and insufficient spatiotemporal perception of traditional methods, enabling accurate understanding of users' multimodal demands and dynamic adaptation in route planning.

The core contribution of this paper lies in transcending the traditional multimodal fusion paradigm and proposing two mechanisms with theoretical innovation.

- 1 Heterogeneous graph construction method based on timestamp alignment. Unlike traditional multimodal alignment methods, this paper innovatively constructs the node and edge structures of heterogeneous graphs based on timestamps, achieving precise alignment and fusion of cross-modal spatiotemporal features. This method not only considers the semantic correlation between modalities, but more importantly, captures the dynamic evolution patterns of multi-modal data in the continuous time dimension.
- 2 Convolution layer dynamic weighting mechanism based on Pearson correlation coefficient. Different from the conventional processing methods of graph convolution networks, this paper proposes an adaptive weight allocation strategy based on correlation coefficients. This strategy dynamically adjusts the contribution degree of features from different time spans by calculating the statistical correlation between each convolutional layer's features and the overall feature distribution, providing a new theoretical framework for cross-modal spatiotemporal feature fusion.

Research on general tourism intent understanding is divided into unimodal-based tourism intent recognition and multimodal-based tourism intent recognition. Unimodal-based intent recognition research mainly involves text, image, and voice modalities. Tourism intent recognition is a technology that extracts travel-related decision-making intentions such as travel demands, destination preferences, and consumption tendencies from users' modal data. Due to the differences in data form, expression characteristics, and information carriers among the three single modalities – text, image, and voice, distinct recognition method systems have been formed. Text can directly express the core tourism intent through sentences, without the need for complex feature reasoning, making it the easiest modality to extract intentions. Images are the most intuitive and visually expressive modality for tourism intent expression, with visual features as the core information carrier. Intentions are mostly implicitly expressed through visual elements without direct language expression, belonging to implicit intent expression. Voice is the most natural and immediate modality for tourism intent

expression. The core information carrier is the acoustic and linguistic features of voice, which is a fusion of acoustic features and linguistic features. Intentions are expressed with both the explicit nature of spoken language and the implicit nature of acoustic features. Huang et al. (2025) used bidirectional encoder representations from transformer (BERT) for textual feature representation of tourist travelogues and employed long short-term memory (LSTM) to obtain tourists' intents, achieving an identification accuracy rate of only 78.9%. Li et al. (2018) developed a bidirectional convolutional neural network-recurrent neural network (CNN-RNN) model integrated with an attention mechanism to improve the performance of tourism intent recognition. Images, as important carriers of information expression, also convey many intent-related details. Image-based recognition methods typically focus on shallow features such as texture and colour of identified targets to achieve feature extraction for target objects. Sharma et al. (2022) used images of different travel-themed comments shared by tourists and applied an SVM classifier to extract image features for intent recognition. Kwon and Yu (2024) compared local regions with emotional information and the entire image using CNN to obtain emotion scores, enabling prediction of tourists' emotions. Qi et al. (2019) used transfer learning to pre-train on the AlexNet model and then fine-tune it on the dataset to obtain the weight parameters, achieving tourist intention recognition with an accuracy rate of 81.33%. Zhou and Huang (2024) proposed a convolutional recurrent neural network based on three-dimensional attention for identifying audio features and tourist intents. Pantano and Styliadis (2021) injected local voice features, location features, etc., into a convolutional multi-channel neural network for tourist intent recognition, improving the identification accuracy.

Owing to the inherent complexity and variability of multimodal data, most approaches in intent recognition rely on pre-trained models to capture characteristics from individual modalities in the feature extraction phase. Pre-trained models are trained on large-scale multimodal datasets before specific tasks, making their performance often better than models trained from scratch for those tasks. By learning from large-scale data, pre-trained models can effectively extract features from different modalities (Si et al., 2024). Calderón-Fajardo et al. (2025) proposed a pre-training model called the multi-interactive memory network (MIMN), which adopts an attention mechanism to guide the model in generating attention vectors for text and images, improving the accuracy of tourism intent recognition. Chen et al. (2025) proposed a cross-modal modulation and gating network model that uses an improved Transformer as the cross-modal modulation module to fuse heterogeneous modal information. Gupta et al. (2020) proposed an attention-based temporal convolutional network to capture unimodal characteristics, enhancing characteristic representation ability, and then introduced a multi-level characteristic integration model to improve the effectiveness of multimodal fusion.

Currently, the vigorous development of tourism has made users' expectations for travel experiences increasingly personalised and specific. As a result, the demand for tourist route planning has also become complex and diverse. However, most existing tour route planning applications are based on single objectives, making it difficult to meet users' diversified and clear requirements. The essence of travel route planning can be abstracted as the modelling and solving of a weighted directed graph: scenic spots and destinations are abstracted as the nodes of the graph, the transportation and movement methods between scenic spots are abstracted as the edges of the graph, and the optimisation objective is abstracted as the weight of the edges. The route planning problem is thus transformed into finding the node traversal path with the optimal weight in this weighted graph, under the condition of meeting certain constraints. The existing single-objective algorithms all revolve around this abstract model. According to the core requirements of route planning, they are divided into three categories: point-to-point optimal path algorithms, multi-scenic spot sequential planning algorithms, and scenic spot combination selection algorithms. They respectively adapt to three single-objective scenarios movement between two points, fixed sequential planning of scenic spots, and sequential planning of scenic spots without fixed locations. Zhang and Sun (2024) combined the k-means clustering method with travelling salesman problem (TSP) to generate travel itineraries. In this process, distance and travel time were considered as key factors for optimal routes. Absalamova et al. (2025) suggested a route planning model aware of travel intent. The model uses a neural network to learn travel intentions and designs a tour route planning algorithm based on heuristic search. Stoyanova-Doycheva et al. (2020) proposed an n-day tour route generation method using q-learning techniques. It not only considers time constraints but also integrates attributes such as ratings, travel time, and cost to meet users' personalised needs and preferences. Beed et al. (2020) suggested a tourist route planning strategy in light of multi-objective optimisation and particle swarm algorithms. This strategy sets accuracy, popularity, and diversity as planning objectives. He (2023) proposed a tour route planning model based on the ant colony optimisation algorithm to generate personalised travel routes that match user preferences.

Analysis of current research on travel intent recognition and route planning reveals that traditional methods for understanding travel intent and planning routes often rely on single-modal data or simple rules, making them ill-equipped to handle complex and dynamic travel scenarios and diverse user demands. To address this, this paper suggests a general approach for travel intent understanding and route planning based on multimodal spatio-temporal pre-trained models. The major contributions of this method lie in four aspects.

- 1 A fine-grained feature extraction scheme was designed to address the distinct characteristics of multimodal tourism data. For textual information, character-level and word-level features were extracted using the robustly optimised BERT pretraining approach

(RoBERTa) model, with a feature alignment module fusing the two sets of information. For video information, feature extraction was performed using a visual transformer. For audio information, feature extraction was conducted via the SpeechT5 encoder.

- 2 Introducing bidirectional LSTM to capture temporal characteristics from each modality's time series. Incorporating self-attention mechanisms to enhance feature weighting within each modality, the automatically filtered feature matrix is fed into graph convolutional neural networks for semantic interaction. Feature aggregation is performed based on timestamps, calculating correlation coefficients in the aggregation layer to obtain fused joint features, thereby enabling cross-modal spatial interaction. A fully connected neural network completes the identification of general tourism intent.
- 3 In view of the problem that the existing tourism route planning methods generally only consider a single demand and pursue the optimum of a single goal, and it is difficult to achieve multi-objective balanced optimisation, this paper proposes a multi-objective tourism route planning method that integrates travel intention, travel cost and redundant time. This method introduces strategies such as local heuristic search and cross-mutation, comprehensively considering multiple objectives including the user's travel intention, redundant travel time and total travel cost, and thereby generating a set of multi-objective balanced and optimal route schemes.
- 4 Simulation experiments were conducted on real datasets. The results indicate that the intention recognition accuracy of the suggested method has increased by at least 3.47%, and the anti-generation distance of route planning has decreased by at least 0.7. There have been significant improvements in both intention recognition accuracy and route planning efficiency, providing a new idea for intelligent tourism services.

2 Relevant theoretical analysis

2.1 Pre-trained model

Pre-trained models have become a fundamental tool for multimodal understanding by learning general features and patterns on large general-purpose datasets, and then fine-tuning to adapt to downstream tasks (Wang, et al., 2023). The pre-trained model adopted in this paper is designed to extract high-quality features for each modality. In multimodal intent recognition, pre-trained models can comprehensively extract features from various modalities, and their performance is superior to that of traditional models. This is mainly due to the fact that pre-trained models, through unsupervised/self-supervised learning with massive general data, can capture the common, essential and transferable feature representations in modal data. After

undergoing small-sample fine-tuning for downstream tasks, they can precisely adapt to specific tasks such as multimodal intent recognition. The collaborative effect of multiple factors such as data scale and distribution, training paradigm and network architecture enables pre-trained models to learn more essential and comprehensive modal features, thereby better meeting the core requirements of cross-modal feature fusion in multimodal intent recognition.

Pre-trained models have become a fundamental tool for multimodal understanding by learning general features and patterns on large general datasets and then fine-tuning to adapt to downstream tasks (Marks et al., 2025). The pre-trained model adopted in this paper aims to extract high-quality features for each modality.

- 1 The text pre-trained model acquires the context information of words, sentences, and paragraphs through large-scale unsupervised learning. BERT and RoBERTa learn rich language features and semantic information by pre-training on large datasets. BERT uses a bidirectional training method, while RoBERTa further improves the model's performance by optimising the pre-training process and strategies of BERT.
- 2 The image pre-trained model ViT [25]. This model uses an architecture similar to that of Transformers in the field of natural language processing to learn global information of images.
- 3 The audio SpeechT5 model. This model adopts the Transformer architecture and processes both speech and text modalities through a shared encoder-decoder structure (Perezhohin et al., 2024). In the pre-training stage, the self-attention and self-encoding objectives are used to learn speech-text alignment, and joint training with masked language modelling (MLM) and connection timing classification (CTC) enables the model to have zero-shot cross-language transfer capabilities.

Pre-trained models are trained on massive general data and can learn common features across tasks and domains. Text pre-training models (such as RoBERTa) use bidirectional Transformer encoding and dynamic masking strategies to accurately capture the semantic context of the text, and are suitable for deep semantic mining of tourism texts. The image pre-training model ViT divides images into blocks and inputs them to the Transformer, using the self-attention mechanism to extract global spatial correlation features, which helps in understanding the scene intent of tourism images. The audio pre-training model SpeechT5 combines MLM and connectionist temporal classification for joint training, taking into account both acoustic features and semantic understanding, enabling end-to-end tourism intent recognition and simplifying the model structure.

2.2 Multimodal fusion technology

Multimodal fusion technology is a method that integrates information from different modalities. Through specific

fusion strategies, this technology extracts richer feature representations, thereby enhancing the performance of the model in various tasks. Currently, the mainstream multimodal fusion methods can be classified into early fusion, late fusion, and hybrid fusion (Castro et al., 2020).

- 1 Early fusion. After extracting the features of each modality, they are mapped to a shared semantic space and fused into a single feature vector, which is then input into the classifier for intent recognition. The advantage is that it captures modality correlations and only requires one classifier, saving computing power. The limitation is that it needs to handle the differences of different modality features and align the feature spaces.
- 2 Late fusion. After extracting and classifying the independent features of each modality, the fusion decision results are output as the final result. The advantage is that it adapts to the feature extraction and classification of each modality's characteristics, achieving local optimality. The limitation is that it ignores the correlations between modalities and requires training multiple classifiers, which is time-consuming and results in low model efficiency.
- 3 Hybrid fusion. By combining early and late fusion, the fusion strategy is adapted according to the modality characteristics. The advantage is that it combines the strengths of both, resulting in better performance. The challenge is to design decision rules that balance the modality relationships and solve the problem of modelling dynamic multi-modal information.

3 Generalised tourism intent recognition based on multimodal spatiotemporal pre-trained models

3.1 Feature extraction from tourism databased on multimodal pre-trained models

Generalised tourism intent recognition involves multimodal data covering key visual, textual, and audio information. Through deep mining of these multimodal data, generalised tourism intents can be understood and recognised. A generalised tourism intent recognition model based on multimodal spatiotemporal pre-training models (MSPM) is proposed. The structure of the MSPM model is shown in Figure 1. First, features are extracted from multimodal tourist data using multimodal pre-training models. A bidirectional LSTM (BiLSTM) is introduced to extract temporal sequence features across each modality. The self-attention mechanism is added to reinforce the weight assignment of intra-modal features, and the automatically selected feature matrix is input into a GCN for semantic interaction. Then, based on timestamps, feature aggregation is performed, the correlation coefficients of the aggregation layer are calculated to obtain fused joint features, achieving cross-modal spatial interaction. Finally, generalised tourism

intent recognition is completed through a fully connected neural network.

- 1 Textual feature extraction. Tourism textual data includes dialogue content, travel planning requests, travel journals, accommodation reviews, etc. This paper uses the RoBERTa pre-training model to extract text features. RoBERTa improves upon the BERT model by adopting a dynamic masking strategy, thereby enhancing the performance of the pre-training model. First, touristic texts Text are classified into i subwords, with each subword represented as an embedding vector S_i composed of the superposition of character vectors P_i , sentence vectors E_i , and positional vectors C_i , as shown below.

$$E_i = C_i + S_i + P_i \quad (1)$$

E_i is multiplied respectively by 12 self-attention head weight matrices, each with initially different weights, to obtain the Q, K, V matrices needed for computing attention values, as shown below. In this context, E_i represents the input vector of the model, and W_i^Q, W_i^K, W_i^V represent the three weight matrices of the self-attention head.

$$\begin{cases} Q_i = E_i \times W_i^Q \\ K_i = E_i \times W_i^K \\ V_i = E_i \times W_i^V \end{cases} \quad (2)$$

Concatenate all attention head output vectors and multiply by a weight matrix W^0 to gain the final output value Z of the multi-head attention layer, as shown in equation (3). The calculation of the attention heads is presented in equation (4), where d_k stands for the vector dimension of Q and K .

$$Z = \text{Concat}[M_1, \dots, M_i] W^0 \quad (3)$$

$$M_i = \text{Attention}(Q_i, K_i, V_i) = \text{softmax}\left(\frac{Q_i K_i^T}{\sqrt{d_k}}\right) V_i \quad (4)$$

Further through residual connections and fully linked layers, map Z into a high-dimensional space and use ReLU to restore it to the original dimensions. Then, pass the result through another residual connection and fully linked layer to the next encoder. At the same time, take the output of the last layer as the RoBERTa output, achieving the final textual feature representation, as shown in equation (5).

$$T = \text{RoBERTa}(H) = [t_1, t_2, \dots, t_n] \quad (5)$$

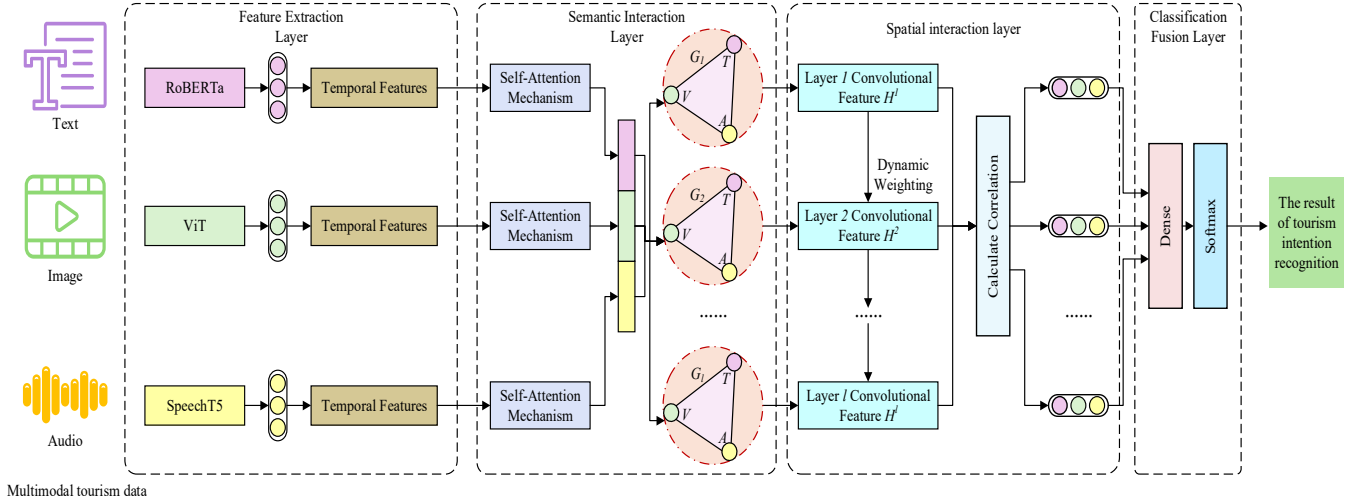
where $H = [h_1, h_2, \dots, h_n]$ is the text character sequence, and h_i represents the contextual vector of the n th character.

- 2 Visual feature extraction. Photographs taken by tourists during their travels and images shared on social media belong to the visual modality of tourism data. This paper uses ViT for feature extraction from image data. First, the image I is processed through a vision transformer to obtain the feature representation $Z = \text{ViT}(I)$. Next, Z passes through the standard decoder, simple decoder, and bottom-up decoder separately. The standard decoder performs upsampling via transposed convolution and generates key points through prediction layers. Transposed convolution reverses the process of convolution and effectively restores spatial resolution. Its output is $C = \text{ConvTrans}(Z)$. In the bottom-up decoder, a branch structure is used to independently predict each key point. The bottom-up decoder can predict the coordinates of different key points in parallel, without relying on traditional sequential structures; its output is $BU = \text{BottomUp}(Z)$. Finally, combine the outputs from the above decoders to obtain the final visual feature representation denoted as V .
- 3 Audio feature extraction. Speech data generated during tourism are rich and diverse, covering interaction scenarios between users and tourist environments, service staff, travel companions, and intelligent devices. This paper uses SpeechT5 (Bolozbekova et al., 2025) to extract features such as intonation, stress, and rhythm. SpeechT5 is a pre-trained model in light of an encoder-decoder structure. First, the raw speech signal U is converted into a spectrogram through short-time Fourier transform, as shown in equation (6), where $STFT$ is the Fourier transform and t is time. Next, apply Mel filter banks Φ to map the spectrogram into the Mel spectral space, resulting in an 80-dimensional Mel spectrogram, as shown in equation (7). Further, a convolutional layer is applied on the Mel spectrogram to extract local features, as shown in equation (8), where W is the convolution kernel, b is bias, $*$ represents the convolution operation, and σ is the activation function. Finally, during the convolution process, downsampling is achieved through stride operations, compressing the feature dimensions to a size suitable for sequence modelling, thereby reducing computational burden. The generated audio features are denoted as A .

$$S = |STFT\{V(t)\}| \quad (6)$$

$$M(t) = \Phi \cdot \log(S + \varepsilon) \quad (7)$$

$$F(t) = \sigma(W^* M(t) + b) \quad (8)$$

Figure 1 The structure of the MSPM model (see online version for colours)

3.2 Temporal feature representation based on BiLSTM

Since the alignment task of multi-modal high-dimensional features is usually completed at the word level, the features extracted from each modality are averaged along the dimension of words as the corresponding high-dimensional features for data alignment, denoted as d_m . All sentence-level features are stored in matrix $X^m = [x_1^m, x_2^m, \dots, x_t^m, \dots, x_L^m]$, where $m \in \{T, V, A\}$ represents modalities, T denotes text, V denotes image, and A denotes audio.

Multi-modal tourism data inherently possess strong temporal characteristics. After storing the features obtained from each feature extractor into X^m , BiLSTM is used to explore temporal sequence features and dependencies across modalities. First, LSTM captures long-range semantic relationships and addresses gradient vanishing or exploding issues during backpropagation. The computational process involves input vectors sequentially passing through the input gate, forget gate, and output gate, followed by a comparison with the internal state c_{t-1}^m of the memory cell from the previous moment to produce output features h_t^m , as shown in equation (9) to equation (14).

$$i_t^m = \sigma(U_i^m x_t^m + W_i^m h_{t-1}^m + b_i^m) \quad (9)$$

$$f_t^m = \sigma(U_f^m x_t^m + W_f^m h_{t-1}^m + b_f^m) \quad (10)$$

$$o_t^m = \sigma(U_o^m x_t^m + W_o^m h_{t-1}^m + b_o^m) \quad (11)$$

$$\tilde{C}_t^m = \delta(U_c^m x_t^m + W_c^m h_{t-1}^m + b_c^m) \quad (12)$$

$$C_t^m = f_t^m \circ c_{t-1}^m + i_t^m \circ \tilde{C}_t^m \quad (13)$$

$$h_t^m = o_t^m \circ \delta(C_t^m) \quad (14)$$

where i_t^m , f_t^m , and o_t^m represent respectively the input gate, update gate, and output gate for the m^{th} modality at the

t^{th} utterance segment, x_t is the input vector of the t^{th} utterance segment, h_{t-1}^m represents the output response from the previous utterance for the m^{th} modality. U , W , and b are weight matrices and bias vectors respectively, σ denotes Sigmoid function, $\tilde{C}_t^m$ indicates candidate state, δ is tanh function, and $\circ$ represents Hadamard product.

LSTM can capture the contextual relationships within modalities, and BiLSTM can obtain context representations of semantic relationships across modalities, as shown in equation (15) to equation (17).

$$h_t^m = LSTM(x_t^m) \quad (15)$$

$$\bar{h}_t^m = LSTM(\bar{x}_t^m) \quad (16)$$

$$z_t^m = [h_t^m \oplus \bar{h}_t^m] \quad (17)$$

where h_t^m represents forward computation using LSTM, $\bar{h}_t^m$ denotes backward computation with LSTM, $\oplus$ indicates concatenation operation, and z_t^m is the final temporal feature matrix.

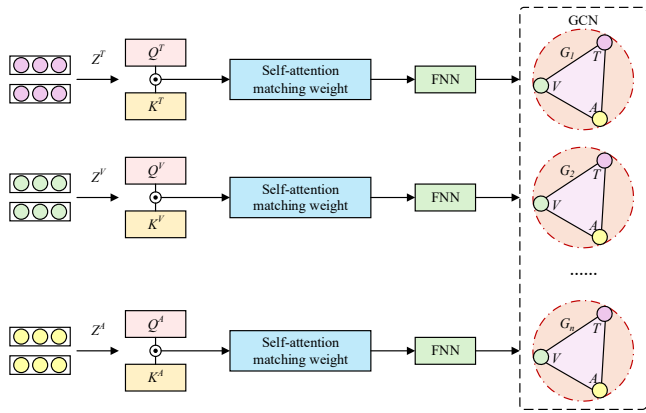
In the multi-modal intent recognition model for tourism data, the semantic interaction layer and the spatial interaction layer are the core levels of collaborative modelling. These two layers conduct double-layer feature mining based on the semantic and spatial attributes of tourism data, ultimately achieving tourism intent recognition. By combining the design logic of BiLSTM in the paper for extracting temporal features of tourism data and the semantic interaction, the spatial interaction layer is not an independent module but rather supplements the modelling of the features of tourism data from the spatial dimension. It forms a triple core connection of dimension complementarity, feature mutual feedback, and reasoning mutual verification between the semantic interaction layer, and the interaction design between the two is highly adapted to the inherent characteristics of tourism data, such as strong

spatiotemporal attributes and tight binding of semantics and space.

3.3 Semantic interaction based on self-attention mechanism

After obtaining context representations within each modality through the time series feature extraction module, to preserve features more relevant to the task in each modality, a self-attention mechanism is added in the semantic interaction module for dynamic weight matching within modalities. The selected feature matrices are then fed into a heterogeneous graph and aggregated based on timestamps, as shown in Figure 2.

Figure 2 Semantic interaction layer (see online version for colours)



In the semantic interaction layer, first, text temporal feature matrix Z^T , image temporal feature matrix Z^V , and audio temporal feature matrix Z^A are respectively input into self-attention mechanism modules to obtain time series data across modalities through key-value queries after internal weight matching. Then a feedforward neural network (FNN) is used to enhance the model performance, and finally the processed vector matrices from the FNN are fed into GCN where text modality T , image modality V , and audio modality A are aligned based on timestamps.

3.4 Spatial feature representation based on graph neural network

At the semantic interaction layer, the weight allocation of semantic elements within a modality and the feature alignment and aggregation based on timestamps between modalities have been achieved. The spatial interaction layer innovatively handles cross-modal spatio-temporal correlations, and its core contributions are manifested in the following two aspects.

- 1 Heterogeneous graph construction method based on timestamp alignment. Traditional multimodal fusion methods usually adopt simple feature concatenation or attention fusion, ignoring the precise alignment in the time dimension. This paper proposes a timestamp-aligned heterogeneous graph construction method, which regards the text, visual, and audio

features at the same timestamp as heterogeneous nodes, connects different time-stamped nodes through time edges, and connects different modal nodes within the same time-stamp through modal edges. This graph structure innovatively encodes two types of relationships.

- 2 Dynamic weighting mechanism of convolutional layer based on Pearson correlation coefficient. After message passing through multiple layers of GCN on the above heterogeneous graph, we obtain the node representations $\{H^{(1)}, H^{(2)}, \dots, H^{(L)}\}$ for each layer.

Traditional methods often use summation, averaging or the output of the last layer as the final representation, assuming that all layers are equally important. This paper proposes that different convolutional layers capture different granularities of information, and their contributions to the final intention recognition should be different. Therefore, this paper introduces a layer importance evaluation mechanism based on statistical correlation. Specifically, first, the feature means of all layers H_{AVG} are calculated as the proxy of global consensus. Then, the Pearson correlation coefficients $\rho(H^{(l)}, H_{AVG})$ between the features $H^{(l)}$ and H_{AVG} of each

layer are calculated. It integrates the traditional statistical metrics used for post-analysis in a forward-looking manner into the adaptive fusion controller in the forward propagation of the model. This not only provides an interpretable decision basis for the feature fusion process of neural networks, but also establishes a new model design idea connecting classical statistics and deep learning. Its value is far greater than manual setting or implicit learning of fixed weights through a fully connected layer as an engineering strategy.

These two mechanisms jointly constitute the core contribution of this paper at the methodological level, going beyond the traditional multimodal pipeline design, and providing a new theoretical framework and solution for addressing the problem of spatio-temporal multimodal fusion.

In this scheme, the modal features of the same timestamp are aggregated as nodes for each layer of graph convolution, and each convolution layer is regarded as a whole. At the same time, the average features of all convolution layers are obtained. Then, the Pearson coefficient is set to dynamically weight different convolution layers to capture the correlations between convolution layers based on the timestamp. Since in the graph classification task, the node feature vectors need to be converted into graph feature vectors, the weighted convolution layer features are transformed into graph feature vectors through the readout function, ultimately achieving spatial interaction. The process of constructing heterogeneous graphs based on timestamp alignment is as follows.

- 1 Node definition. In the graph, each node $v_t^m \in V$ represents a modal-time stamp basic unit, where, $m \in \{T, V, A\}$ stands for text (T), vision (V), or audio (A), and t is the discretised timestamp number. Each node is characterised by the feature vector C of the corresponding modal at time t (initialised). Therefore, a data instance containing L time segments and 3 modalities will be transformed into a heterogeneous graph consisting of $3L$ nodes.
- 2 Edge definition and adjacency relationship construction. This paper defines three types of edges to encode different association priors, collectively forming the adjacency relationship (v_t^m, v_{t+1}^m) of the graph.
Firstly, connect the nodes $(v_{t_i}^{m_i}, v_{t_j}^{m_j})$ of the same modality m at adjacent timestamps to capture the state evolution and continuity of each modality itself. Secondly, connect the nodes of different modalities at the same timestamp t to model the complementary descriptions of different modalities for the same scene or event at the same moment. To enhance the capture of long-range dependencies, we also add learnable global attention edges between all nodes.
- 3 Timestamp alignment mechanism. Timestamp alignment forms the basis of graph construction. For text modalities, we take discourse segments as the basic units, and each segment is assigned an ordered timestamp. For images and audio, they are allocated to corresponding timestamps based on their correspondence with the text description. Data from all modalities at the same timestamp t are regarded as descriptions of the same moment of tourism experience or user state, thereby forming a cross-modal alignment unit. This alignment method ensures that heterogeneous graphs can accurately reflect the coexistence and evolution relationship of multi-modal data on the time axis.

After completing the construction of the aforementioned heterogeneous graph, this paper feeds it into a multi-layer graph convolutional network (GCN) for spatial interaction. To optimise the fusion effect, this paper further proposes a dynamic layer weighting strategy based on the Pearson correlation coefficient, which adaptively fuses the output features $H^{(l)}$ of each layer of the GCN to obtain a more discriminative graph-level representation H_G . In the GCN, a $K = 2$ -layer graph convolution is adopted. The input node feature dimension is $d_{in} = 768$, the hidden layer dimension is $d_{hidden} = 256$, and the output dimension is $d_{out} = 128$.

In GCN, the features aggregated for each convolutional layer are usually directly summed. However, in cross-modal sentiment analysis problems, information corresponding to different timestamps has varying importance levels. To address this issue, a weighted feature fusion method is adopted to assign weights to each convolutional layer. Equation (18) represents $H^{(l)}$ all node information obtained

at the l^{th} layer, and the mean of features across N layers of convolution is represented by equation (19).

$$H^{(l)} = \text{Relu} \left(\tilde{D}^{-\frac{1}{2}} \tilde{A} \tilde{D} H^{(l-1)} W^{(l)} \right) \quad (18)$$

$$H_{AVG} = \frac{1}{N} \sum_{l=1}^N H^{(l)} \quad (19)$$

where $\tilde{D}$ is the degree matrix of the self-connection adjacency matrix, $H^{(l-1)}$ is the input features at layer l , $H^{(l)}$ is the output features at layer l , $W^{(l)}$ is the linear transformation matrix, and $\tilde{A}$ is the self-connection adjacency matrix.

The Pearson correlation coefficient $\rho_{(H^{(l)}, H_{AVG})}$ is calculated by using the feature mean of all nodes in each layer and the N -layer convolutional feature mean to obtain their correlation. The higher the $\rho_{(H^{(l)}, H_{AVG})}$, the higher the similarity, indicating that the performance of the convolutional layer is better; therefore, a correspondingly greater weight should be assigned. The similarity relationship between convolutional layers can thus be expressed as follows, where $\text{cov}(H^{(l)}, H_{AVG})$ represents the covariance, which is mainly contributed by the core semantic components of $H^{(l)}$ and H_{AVG} . Therefore, the higher the correlation, the higher the proportion of the core semantic component of $H^{(l)}$, and the greater its semantic contribution to the intention recognition. This theoretical logic establishes a direct quantitative correlation between statistical correlation and semantic contribution, proving the theoretical rationality of this weighting mechanism.

$$\rho_{(H^{(l)}, H_{AVG})} = \frac{\text{cov}(H^{(l)}, H_{AVG})}{\sigma_{H^{(l)}} \sigma_{H_{AVG}}} = \frac{E(H^{(l)}, H_{AVG}) - E(H^{(l)})E(H_{AVG})}{\sqrt{E(H^{(l)2}) - E^2(H_{AVG})}} \quad (20)$$

By obtaining the correlation between each convolutional layer and N -layer convolutions through the correlation coefficient, the weights for each convolutional layer are determined. Then, by performing calculations, the final representation of each convolutional layer is obtained, as shown in equation (21).

$$H_F^{(l)} = H^{(l)} \cdot \rho_{(H^{(l)}, H_{AVG})} \cdot \alpha \quad (21)$$

where $H_F^{(l)}$ represents all node features at layer l 's convolution, $\rho_{(H^{(l)}, H_{AVG})}$ represents the Pearson coefficient, and α is a trainable global scale factor. α , along with all other weight parameters in the model, is updated through the backpropagation algorithm and gradient descent. Its gradient is derived from the final intent classification loss function, thus the optimisation objective is to make the entire weighted fusion mechanism most beneficial for the downstream classification task. This paper does not impose any hard boundaries on α . It is a free parameter that is initialised to 1 in the early training stage. This enables the

model to start from an unscaled initial state and freely learn an appropriate scaling amplitude based on the training data, thereby maximising the utilisation of the relative importance information provided by the Pearson correlation coefficient.

Since in graph classification tasks all node feature vectors need to be converted into graph feature vectors for representation, a readout function is used to aggregate node features and obtain feature representations based on the graph convolutional layer. $H_G^{(l)}$ represents the graph feature of each layer, as shown in equation (22).

$$H_G^{(l)} = \text{readout}(H_F^{(l)}) \quad (22)$$

3.5 General travel intent recognition results output

After spatial interaction, obtained graph features need to be fused for identification. The fusion identification layer consists a fully connected neural network and a Softmax function.

After spatial interaction, each layer's graph features H_G are acquired. To adequately fuse these graph features and obtain a final full-dynamic fused feature representation, the paper adopts a joint architecture from multimodal fusion frameworks for feature fusion and uses a fully connected neural network as the shared semantic subspace in the joint architecture. Map H_G to the shared semantic subspace. Perform linear transformation of input data within the shared semantic subspace. Then apply the activation function $\tanh$ to perform nonlinear processing on results from the fully connected neural network, obtaining an output vector H_s , and acquiring full-dynamic fused features, as shown in Equation (23), where H_s represents the full-dynamic fused feature obtained after going through a fully connected layer, and W_s, b_s represent weights and bias.

$$H_s = \tanh(W_s H_G + b_s) \quad (23)$$

Finally, use a Softmax function to perform general tourism intent classification identification on the full-dynamic fused features H_s , as shown in equation (24), where y represents the predicted probability distribution of tourism intents.

$$y = \text{Softmax}(H_s) \quad (24)$$

3.6 Comparative analysis with existing multimodal fusion methods

The MSPM framework proposed in this paper has fundamental differences from the current mainstream multimodal fusion methods in several key design concepts. This section compares MSPM with various methods to clearly elaborate on its unique advantages and innovative positioning.

- 1 Compared with transformer-based multimodal fusion methods. Traditional transformer-based multimodal fusion methods mainly rely on self-attention and cross-attention mechanisms for global feature interaction in the implicit space. Although these

methods can capture long-range dependencies, their attention weight distribution is usually data-driven and lacks clear structural constraints. The core difference of MSPM lies in the introduction of explicit spatio-temporal structured priors. Through timestamp-aligned heterogeneous graph construction, the temporal co-occurrence relationships and modal complementary relationships between multimodal data are explicitly encoded as edges in the graph structure. This design enables the model to perform relationship reasoning directly on the graph rather than merely aggregating features in the feature space. In the tourism intent understanding scenario, this explicit spatio-temporal relationship modelling can more accurately capture the fine-grained correspondence between users' visual attention, voice descriptions, and text requirements at specific time points.

- 2 Compared with collaborative attention/cross-modal attention methods. Collaborative attention and cross-modal attention mechanisms achieve bidirectional or unidirectional feature re-calibration by calculating attention weights between modalities. These methods are at the feature level of soft alignment. MSPM combines structured hard alignment with statistical soft fusion. Firstly, timestamp alignment provides hard alignment constraints for cross-modal elements, ensuring consistency of spatio-temporal context. On this basis, through a dynamic weighting mechanism based on Pearson correlation coefficient, we achieve adaptive fusion of aligned features. This two-stage strategy of first aligning and then optimising fusion is more robust in handling common modal asynchrony or noise issues in tourism scenarios compared to a single attention mechanism.
- 3 Compared with CLIP-style alignment methods. CLIP and its derivative models achieve cross-modal semantic alignment in the latent space through large-scale contrastive learning. The core advantage of these methods lies in their strong zero-shot transfer ability, but their alignment is global and semantic-level. MSPM addresses the dynamic alignment problem in fine-grained spatio-temporal scenarios. The heterogeneous graph of MSPM is aligned along the timestamp axis, achieving this fine-grained, serialised cross-modal instance alignment, which is an important supplement and scenario-specific deepening of the global alignment style of CLIP.

4 Multi-objective tour route planning integrating tourist intentions

During the process of planning a travel route, users often have to make choices among multiple conflicting goals. These goals are difficult to achieve optimally simultaneously, which significantly increases the complexity of the decision-making process. The Pareto optimal solution set can provide users with a quantitative

trade-off scheme among the goals, helping them make reasonable route planning decisions based on their travel intentions and actual travel conditions (Azarberahman et al., 2025). Considering the multi-constraint, multi-objective, and highly personalised characteristics of travel route planning, this paper proposes a hybrid multi-objective optimisation model. This model integrates the classic strength Pareto evolutionary Algorithm II (SPEA-II) with a local search strategy based on heuristic results. At the same time, to meet the actual needs of the travel scenario, an adaptable heuristic function has been redesigned, and the construction of the fitness function and the population update mechanism of the SPEA-II algorithm have been adjusted specifically.

4.1 Problem description and related definitions

4.1.1 Definition of decision variables

The decision variables of the multi-objective tourism route planning model proposed in this paper focus on the selection of tourist attractions, the sequence of visits, and the route allocation for users' single-day or multi-day travel plans. Considering the actual scenarios of users' travel, three types of core decision variables are clearly defined to avoid ambiguity.

- 1 Binary decision variable x_{ij} . It is used to represent the access relationship between tourist attractions. $x_{ij} \in \{0, 1\}$, where i and j represent different tourist attractions ($i, j \in S$, S is the set of all available attractions, $i \neq j$). If $x_{ij} = 1$, it indicates that the user travels directly from attraction i to attraction j during the trip. If $x_{ij} = 0$, it indicates that the user does not travel directly from attraction i to attraction j .
- 2 Binary decision variable y_{ik} . It is used to represent the affiliation relationship between the scenic spots and the sub-routes. $y_{ik} \in \{0, 1\}$, where k is the sub-route number ($k \in K$, K is the set of sub-routes). If $y_{ik} = 1$, it indicates that the scenic spot i is assigned to the k^{th} sub-route. If $y_{ik} = 0$, it indicates that the scenic spot i is not assigned to the k^{th} sub-route.
- 3 Continuous decision variable t_{ik} . It is used to represent the duration of stay of the user at attraction i (allocated to the k^{th} sub-route). $t_{ik} \geq 0$, whose value range is determined by the recommended visiting duration of attraction i ($t_{i,\min} \leq t_{ik} \leq t_{i,\max}$), $t_{i,\min}$ and $t_{i,\max}$ represent the shortest and longest recommended visiting time of attraction i respectively.

4.1.2 The distinction between the number of routes and individual trips

To clarify the differences between the two and avoid confusion, this article, based on the actual scenarios of tourism route planning, provides the following clear definitions.

- 1 Single itinerary. It refers to the complete travel plan of the user during this tourism activity. It includes the starting and ending points set by the user, the total time budget, the total cost budget, the core travel requirements (determined based on the travel intention), and other global constraints. Essentially, it is the overall goal set by the user for this trip and the boundary conditions.
- 2 Number of routes. It refers to the number of sub-routes divided within the framework of a single itinerary to ensure reasonable time allocation, smooth route connection, and efficient coverage of scenic spots (i.e., the above decision variables $|K|$ and $|K| \geq 1$). It is a local execution unit.

Multiple sub-routes together form a single itinerary. All sub-routes must comply with the global constraints of the single itinerary. Moreover, sub-routes should satisfy the principles of no conflict in connection and no repetition of scenic spots. When $|K| = 1$ is true, a single itinerary is a complete sub-route and does not need to be split.

4.1.3 Optimise the objective function

The multi-objective optimisation goals of this article are designed based on the core needs of users, taking into account three aspects: cost economy, time rationality, and maximum scenic value. The objective function is as follows.

$$\begin{cases} \min F_1 = \sum_{k \in K} \left(\sum_{i \in S} y_{ik} \cdot C_{\text{ticket}}(i) + \sum_{i \in S} \sum_{j \in S} x_{ij} \cdot y_{ik} \cdot y_{jk} \cdot C_{\text{tran}}(i, j) \right) \\ \min F_2 = \sum_{k \in K} \left(\sum_{i \in S} y_{ik} \cdot t_{ik} + \sum_{i \in S} \sum_{j \in S} x_{ij} \cdot y_{ik} \cdot y_{jk} \cdot T_{\text{tran}}(i, j) \right) \\ \min F_3 = -\sum_{k \in K} \sum_{i \in S} y_{ik} \cdot V(i) = -\sum U(i) \end{cases} \quad (25)$$

where F_1 represents the total cost target for a single itinerary, which includes the total cost of entrance tickets for all sub-routes ($C_{\text{ticket}}(i)$ is the entrance ticket cost for attraction i) and the total transportation cost between attractions ($C_{\text{tran}}(i, j)$ is the transportation cost from attraction i to j). F_2 represents the total time target for a single itinerary, which includes the total time spent at attractions and the total transportation time between attractions ($T_{\text{tran}}(i, j)$ is the transportation time from attraction i to j). F_3 is the goal of minimising the loss of attraction value (equivalent to maximising the attraction utility), $V(i)$ is the value score of attraction i , $U(i)$ is the weighted fusion result of user intention and attraction attributes, which comprehensively reflects the compatibility between the attraction and the user's needs, and the mapping formula is designed based on the MSPM intention score, that is, $U(i) = U_0(i) \times (\alpha_1 \cdot I_1 \cdot w_1(i) + \alpha_2 \cdot I_2 \cdot w_2(i) + \alpha_3 \cdot I_3 \cdot (1 - w_3(i)))$, where $U_0(i)$ is the basic utility of the attraction, α_1 , α_2 , α_3 are the intention weights, set according to the confidence level of multimodal intention recognition, $\alpha_1 + \alpha_2 + \alpha_3$. $w_1(i)$, $w_2(i)$ and $w_3(i)$ are the weight of attraction ratings.

4.1.4 Constraint conditions

Taking into account the actual constraints and decision-making logic of the travel route planning, clearly define various constraints, with a focus on detailing the start and end conditions. All constraints must meet both the global requirements of a single itinerary and the local requirements of sub-routes. The specific details are as follows:

- 1 Start and end conditions. Let s_0 be the starting scenic spot of a single itinerary (specified by the user, usually a nearby scenic spot or core landmark near the departure point), and e_0 be the ending scenic spot of a single itinerary (also specified by the user, usually a nearby scenic spot or core landmark near the return destination). The start and end constraints are divided into global constraints and sub-route constraints.

Global start and end constraints. The overall starting point of a single itinerary is s_0 , and the ending point is e_0 . That is, there exists a unique sub-route $k_1 \in K$ that includes e_0 ($y_{s_0 k_1} = 1$), and s_0 is the first visited attraction of this sub-route. There is a unique sub-route $k_2 \in K$ that includes e_0 ($y_{e_0 k_2} = 1$), and e_0 is the last visited attraction of this sub-route. When $|K| = 1$ occurs, $k_1 = k_2$. That is, the same sub-route includes s_0 and e_0 , and s_0 is the starting point and e_0 is the ending point. Sub-route start and end constraints: Each sub-route must have a clear starting and ending attraction, and the starting attraction of a sub-route must be connected with the ending attraction of the previous sub-route, ensuring that there are no connection conflicts between sub-routes.

- 2 Site visit constraints. Each site can be assigned to at most one sub-route to prevent redundant visits. That is,

$$\sum_{k \in K} y_{ik} \leq 1, \forall i \in S. \text{ The sites assigned to a certain sub-route must have both entry and exit points, that is, } \sum_{i \in S} x_{ij} \cdot y_{ik} = \sum_{j \in S} x_{ji} \cdot y_{ik}, \forall j \in S, \forall k \in K.$$

Additionally, at least one site must be visited. This excludes the cases where the starting and ending sites are merely the departure or return points, that is,

$$\sum_{i \in S} \sum_{k \in K} y_{ik} \geq 1.$$

- 3 Time constraints. The duration of each sub-route should not exceed the time budget for that sub-route, i.e.,

$$\sum_{i \in S} y_{ik} \cdot t_{ik} + \sum_{i \in S} \sum_{j \in S} x_{ij} \cdot y_{ik} \cdot T_{\text{tran}}(i, j) \leq T_k, \forall k \in K,$$

where T_k is the time budget for the k^{th} sub-route, and

$\sum_{k \in K} T_k \leq T_{\text{total}}$ (T_{total} is the total time budget for a single trip). The duration of the stay at the attractions falls within the recommended range, i.e., $t_{i, \min} \leq t_{ik} \leq t_{i, \max}$, $\forall i \in S, \forall k \in K$. If $y_{ik} = 0$, then $t_{ik} = 0$, and the constraints are automatically satisfied.

- 4 Cost constraint. The total cost for a single trip shall not exceed the user-defined total cost budget, denoted as

$$\sum_{k \in K} \left(\sum_{i \in S} y_{ik} \cdot C_{\text{ticket}}(i) + \sum_{i \in S} \sum_{j \in S} x_{ij} \cdot y_{ik} \cdot y_{jk} \cdot C_{\text{tran}}(i, j) \right) \leq C_{\text{total}},$$

where C_{total} represents the total cost budget for a single trip.

4.2 Multi-objective tour route planning model

During the process of planning a travel route, users often have to make choices between mutually exclusive goals, as these goals cannot be simultaneously optimised, making the decision-making process complex. However, the Pareto optimal solution set provides users with a set of trade-off solutions among the goals, enabling them to make more reasonable plans based on their intentions and actual circumstances. This paper proposes a hybrid multi-objective optimisation model, which combines the SPEA-II with a local search strategy based on heuristic results. Considering the characteristics of the travel route planning problem, this paper redesigns a specific heuristic function and adjusts the fitness function construction and population update mechanism of the SPEA-II algorithm. Next, this paper will elaborate in detail on the heuristic search strategy and the application of the hybrid multi-objective optimisation model in the route planning process.

This paper employs a local heuristic search strategy to calculate the heuristic results based on the travel cost and the given user time limit. Specifically, during the local search process, a heuristic calculation function is defined for each potential visited attraction, as shown in equation (26), where HR_{ns_i} represents the heuristic result for the next tourist destination ns_i . λ and $(1-\lambda)$ are the weights for cost and time constraints, which can be dynamically adjusted according to the user's preference for the travel intention. $C_{\text{ticket}}(ns_i)$ is the ticket cost for ns_i , $C_{\text{tran}}(s_i, ns_i)$ is the transportation cost from the current location s_i to ns_i , $TR_{ns_i} = T_{\text{remaining}} - (t_{ns_i} + T_{\text{tran}}(s_i, ns_i))$ is the time limit remainder, which is the difference between the remaining time budget for a single trip and the sum of the user's stay time at ns_i and the transportation time from s_i to ns_i . The larger this value is, the more sufficient the user's disposable time will be after visiting this attraction, making it more convenient for subsequent destination selection and route adjustment.

$$HR_{ns_i} = \lambda (C_{\text{ticket}}(ns_i) + C_{\text{tran}}(s_i, ns_i)) + (1-\lambda) TR_{ns_i} \quad (26)$$

In the planning of tourism routes, this paper designs the SPEA-II algorithm, whose core lies in the design of population initialisation, mutation and crossover operations for the route planning problem.

- 1 Generation of the initial population. A hybrid strategy is adopted to ensure the quality and diversity of the initial solution. 60% of the individuals are generated through local search based on heuristic functions, iteratively selecting scenic spots from the starting point and satisfying all constraints; the remaining 40% are

randomly generated under the constraints of time and cost.

- 2 Mutation operation. It aims to reduce time redundancy in the itinerary. By calculating and sorting the remaining time of each sub-route, the sub-routes with more time and those with less time are prioritised for merging, and then local search is used to adjust the sequence of scenic spots and insertions to ensure that the new route satisfies all constraints.

Table 1 Pseudo-code for multi-objective tourism route planning algorithm

Algorithm 1 Hybrid multi-objective optimisation model

Input: Maximum number of iterations T_{max} file size N_A

Output: Route set *RouteList*

```

1  Generate the initial population  $P_1$  of the solution;
2  Generate a random population  $P_2$ ;
3  Combined population  $P = P_1 + P_2$ ; // merge the
   population as the initial population
4  Initialise the external file  $A = \emptyset$ ; //the external file is an
   empty set
5
    $t = 0$ ; //initialisation of the
   number of iterations
6  Repeat
7      Perform non-dominated sorting on the population  $P$ 
   to obtain the non-dominated solution set
8      Assign the non-dominated solution set to the
   external file  $A$ 
9      if the external file is larger than  $N_A$  then
10         Density estimation of the external file  $A$ ;
11         Use truncation selection  $A$ 
12         Remove part of the solution so that the file size
   does not exceed  $N_A$ 
13     end if
14     Initialise the next generation population  $P' \emptyset$ ;
15     for each  $s_i$  in  $P$  then
16         Calculate the fitness value  $Fitness(s_i)$  of
   individual  $s_i$ 
17         Apply the crossover mutation operation to  $s_i$  to
   generate  $s'_i$ 
18         if  $s'_i$  is a non-dominated solution or there is no
   dominated solution in a  $s'_i$  then
19             Add the new individual to the next
   generation population  $P'$ 
20         end if
21     end for
22     Update the population  $P = P'$ ;
23     Number of update iterations  $t = t + 1$ 
24 until  $t = T_{max}$ ; select the solutions that meet the conditions
   from the external file  $A$  as the final route set RouteList?
25 return RouteList

```

- 3 Crossover operation. The core is to inherit the high-quality sub-routes from the parent generation. From each of the two parent generations, the sub-routes with higher evaluation are passed on to the offspring, and the uncovered scenic spots (floating nodes) are reinserted or combined under the constraints to form new sub-routes, thereby inheriting the excellent structure while introducing recombination innovation.

This paper solves the multi-objective tourism route planning problem using the aforementioned improved hybrid multi-objective optimisation model. It strictly follows the aforementioned optimisation formula, decision variables and constraints. The specific process is as follows. The multi-objective tourist route planning model that integrates tourists' intentions is shown in Table 1.

- Step 1 Initialise parameters. Determine the set of selectable attractions S , the starting and ending attractions s_0 and e_0 , set the total time budget T_{total} and total cost budget C_{total} for a single itinerary, and initialise the upper limit of sub-route quantity, population size, and number of iterations, etc.
- Step 2 Generate the initial population. Use a hybrid strategy to generate the initial population. Each individual is a single itinerary that satisfies all constraints and contains several sub-routes.
- Step 3 Fitness evaluation. Based on the optimisation objective functions F_1 , F_2 , and F_3 , combined with the fitness function construction method of the SPEA-II algorithm, evaluate the fitness of each itinerary in the population, and simultaneously verify the satisfaction of the constraint conditions, eliminating invalid itineraries.
- Step 4 Non-dominated sorting. Utilise the non-dominated sorting mechanism of SPEA-II to sort all itineraries in the population, select the non-dominated solution set (Pareto optimal solution candidate set), and store it in the external archive.
- Step 5 Crossover and mutation. For the candidate itineraries in the non-dominated solution set, perform the above crossover and mutation operations to generate new offspring itineraries, ensuring that the offspring itineraries satisfy all constraint conditions.
- Step 6 Population update. Merge the parent population with the offspring population, select the optimal individuals to form the next generation population, and maintain the population size stability.
- Step 7 Termination judgment: if the preset number of iterations is reached, or the non-dominated solution set in the external archive tends to be stable, then terminate the iteration. Otherwise, return to step 3 and continue the iteration.
- Step 8 Output results: after the iteration is terminated, output the non-dominated solution set in the

external archive, which is the multi-objective trade-off travel route plan provided to the user. The user can select the appropriate itinerary based on their own preferences.

5 Experimental results analysis

5.1 Dataset and experimental setup

5.1.1 Experimental dataset

This paper uses multimodal tourism data collected from Ctrip by the reference (Hu and Geng, 2024) as the dataset. Ctrip is China's largest online travel company, providing various travel services and also including travelogues and guides shared after tourists' trips; it has a large amount of real-world multimodal tourist data. The dataset includes 41,528 travelogues published by 4,029 tourists, which contain 19,357 text records, 12,069 images, and 10,102 audio records. In each instance, the text, images and audio are associated by the users at the time of publication, jointly describing the same tourism experience segment. This paper aligns the internal structure of the travel notes to ensure the relevance of the multimodal data. This study defines tourism intentions as five categories: sightseeing tours, leisure vacations, theme tours, adventure exploration, and cultural research and study. Acc-5 and Acc-3 in the experiment respectively refer to the accuracy rate of five-category classification and the accuracy rate of three-category classification.

This paper conducts pre-processing on the data in the dataset. The data is segmented using the jieba tokeniser and a word list is constructed. The maximum length of the input sequence is truncated/padded to 128 words. The word embedding dimension is 768, which is consistent with the RoBERTa model. All images are uniformly scaled to a resolution of 224×224 pixels and data augmentation is performed using random horizontal flipping and colour jittering. The original audio files are uniformly resampled to 16 kHz mono. A 25 ms Hamming window with a step size of 10ms is used for short-time Fourier transform to generate an 80-dimensional Mel spectrogram. The length of the spectrogram is uniformly truncated or padded to 800 frames. Silent segments are removed using energy-based speech activity detection. The pre-processed dataset is divided into training set, test set and validation set in a 6:2:2 ratio.

The geographical distances between the attractions $d(s_i, s_j)$ are obtained by calling the path planning interface of the Gaode Map API. The traffic time T_{tran} is determined based on the real-time driving time estimation provided by this API, and it is assumed that the traffic speed is the free-flow speed of the path. The attraction ticket fees C_{ticket} are sourced from public pricing on platforms such as Ctrip and Meituan. The transportation costs C_{tran} are calculated based on the distance and the estimated unit price of taxis or self-driving in the local area. The candidate attractions (POI) and their descriptions are collected from public

tourism platforms, and the stay time is based on the recommended duration by the platform.

5.1.2 Parameter settings

The hardware environment for this experiment is Windows 10 operating system, Intel Core i7 processor, and 32GB of memory. The programming environment is PyCharm 2019.3.3 and Python 3.7. Additionally, the TensorFlow-1.15.3 CPU version is used to run the word embedding code. In the experiment, the learning rate is set to 0.001, the weight decay is $1e-5$, the embedding dimension is 32, the hidden layer neurons are 128, the number of attention heads is 4, the batch size is 200, and the optimiser is Adam. To reduce the influence of randomness, all experiments use a fixed random seed for initialisation and are run independently 5 times. For fair comparison, all comparison algorithms are implemented in Python and run in the same hardware environment. The key parameter settings are as follows.

- HAIE (Absalamova et al., 2025). The initial solution proportion of the local search is 30%, the maximum number of iterations is 1,000, and the neighbourhood is defined as swapping the positions of any two attractions.
- MOTPSO (Beed et al., 2020). The population size is 100, the inertia weight decreases linearly from 0.9 to 0.4, the learning factor $c_1 = c_2 = 2$, and the maximum number of iterations is 500.
- RPMACO (He, 2023). The number of ants is 50, the evaporation factor $\rho = 0.1$ of pheromones, the importance of pheromones $\alpha = 1$, the importance of heuristic information $\beta = 2$, and the maximum number of iterations is 500.
- SPGA (the method suggested in this paper). The population size is 100, the size of the external archive is 100, the crossover probability $p_c = 0.8$, the mutation probability $p_m = 0.2$, and the maximum number of evolutionary generations is 500. Local heuristic search is used to initialise 30% of the population. Stopping criterion. All algorithms adopt the same dual stopping criterion, either reaching the maximum number of iterations, or the Pareto front improving by less than 1% in 50 consecutive generations.

5.2 Performance evaluation of tour route planning algorithms

To verify the performance of the tourism route planning algorithm, this paper constructs a route planning simulation environment based on the spatial and temporal attributes of the attractions extracted from the dataset and the user intention labels. The parameters such as the distance between attractions, travel time, and cost are obtained from public map APIs and service data of Ctrip platform and normalised. All algorithms are implemented in Python and run in the same hardware environment. Each route planning

case is independently run 30 times under randomly generated different user constraints to calculate the average performance.

Figure 3 shows runtime samples executed at different heuristic H levels, using local search heuristics for heuristic filling in 10%, 30%, and 60% initial candidate routes. In the figure, H represents the percentage of local search heuristics used in the initial population by the hybrid algorithm. When the proportion of heuristic results is 30% ($H = 0.3$), the model converges to a better solution than at 10% and 60%. For the remaining part of the population, random filling is adopted to achieve diversity. This hybrid strategy combines the advantages of heuristic search and random search; it can not only use heuristic information to guide the search direction but also maintain population diversity and avoid premature convergence.

Figure 3 Runtime samples executed at different heuristic H levels (see online version for colours)

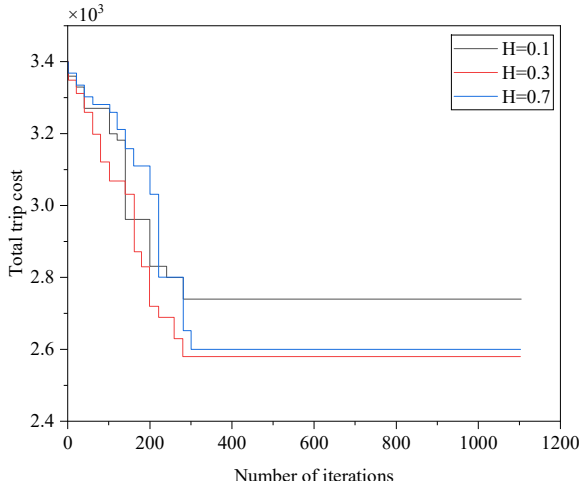
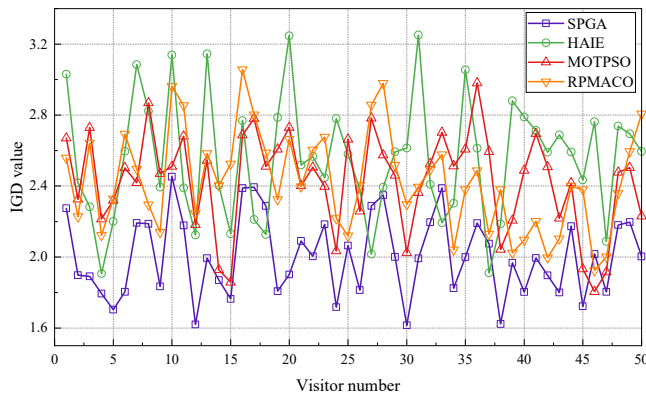


Figure 4 IGD comparison of different tourism route planning methods (see online version for colours)



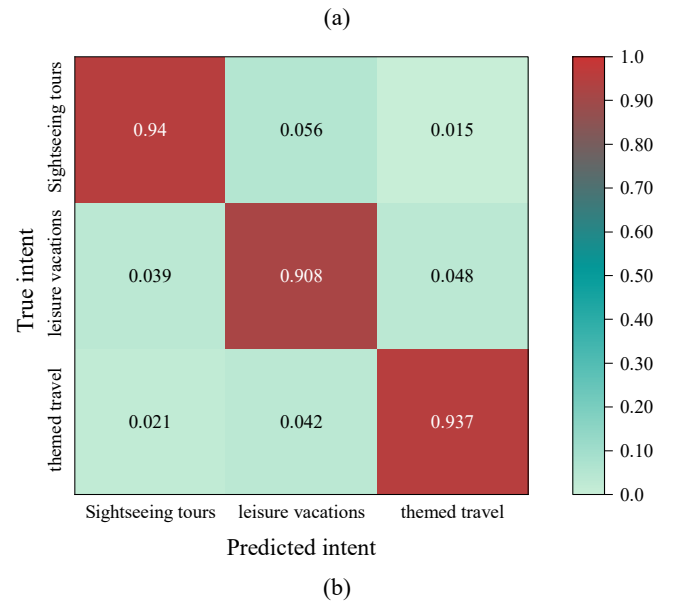
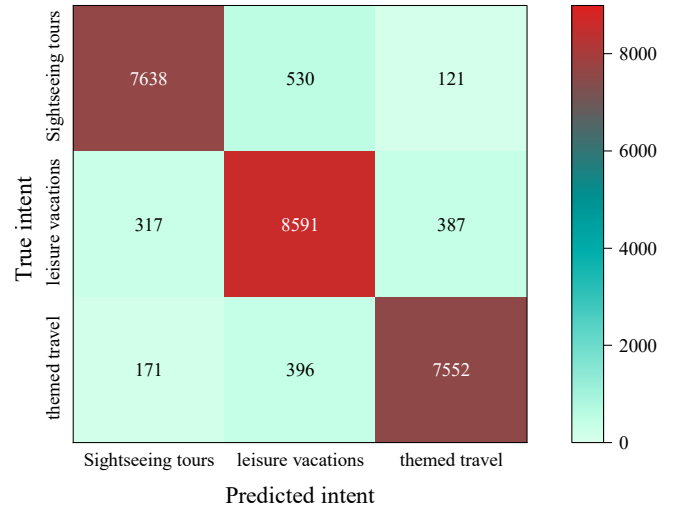
This article compares the designed algorithm (SPGA) with HAIE, MOTPSO, and RPMACO. Currently, the most widely used tool for evaluating multi-objective optimisation results is inverted generational distance (IGD). The smaller the IGD value, the better the algorithm's performance. To reduce random errors, each tourist sample undergoes 30 operations under each method, and the average of the IGD values obtained from these 30 iterations is taken as the final

result. The experimental results are compared in Figure 4. The results show that for 50 groups of tourist samples, the overall IGD obtained by the method used in this paper is smaller than the other three methods, indicating that the method designed for the research problem significantly outperforms the other three methods.

5.3 Performance analysis of travel intent recognition

The confusion matrix of the proposed tourist intent recognition model MSPM is shown in Figure 5. Its diagonal represents the recognition probability for each of the three types of tourist intentions, with accuracy exceeding 90% in sightseeing tours, leisure vacations, and themed tourism scenarios, reflecting that the performance of the MSPM model in recognising tourist intents is relatively strong.

Figure 5 The confusion matrix of the MSPM model, (a) confusion matrix based on sample size, (b) probability-based confusion matrix (see online version for colours)



To further verify the recognition performance of the proposed model MSPM, this paper selects five-class

accuracy (Acc-5), three-class classification accuracy Acc-3 (Acc-3), three-class classification recall, three-class classification F1 score, and AUC value as evaluation metrics for tourist intent recognition. Acc-5 represents the overall classification accuracy of the model across all 5 tourism intent categories, reflecting the model's performance in the complete multi-classification task. To compare with some baseline studies and analyse the model's performance in the core scenarios, this paper merged some semantically similar or less sampled categories, simplified the 5 categories into 3 core categories, and calculated the accuracy rate Acc-3 on this simplified classification task. The baseline methods include CNN-RNN (Li, et al., 2018), MSTL (Qi, et al., 2019), MSCNN (Pantano and Styliadis, 2021), MIMN (Calderón-Fajardo et al., 2025), and AMTCNN (Gupta et al., 2020). The comparison of various recognition performance indicators across different methods is shown in Table 2. Through the experiments, it can be found that the Acc-5 value of MSPM is 87.39%, which improves by 2.22% compared to the second-best method AMTCNN, indicating that the proposed method significantly outperforms other methods in multi-class tourist intent recognition tasks. The Acc-3 value of MSPM is 93.51%, improving from the Acc-3 value of the second-best method AMTCNN (90.04%) by 3.47%, indicating that the proposed method also has a significant advantage in three-class classification tasks. The Recall and F1 score of the proposed method are 92.63% and 94.18%, respectively, which are at least improved by 4.48% and 2.66% compared to the other five models, further validating the superiority of the proposed method in general tourist intent recognition tasks. By comparing the prediction accuracy metric AUC, the AUC value of MSPM is 0.9885, which increases by 18.85%, 14.29%, 9.18%, 7.36%, and 2.61% compared to CNN-RNN, MSTL, MSCNN, MIMN, and AMTCNN respectively. The proposed method enhances semantic interaction and spatial interaction based on temporal information, while also considering the correlation between feature-specific features and overall features, thus improving the accuracy of tourist intent recognition.

Table 2 Comparison of various recognition performance indicators of different methods

Model	Acc-5/%	Acc-3/%	Recall/%	F1/%	AUC
CNN-RNN	71.52	79.08	78.53	81.16	0.8317
MSTL	73.08	81.33	82.47	84.68	0.8649
MSCNN	78.63	84.91	83.75	87.43	0.9054
MIMN	81.04	88.26	86.21	89.66	0.9207
AMTCNN	85.17	90.04	88.15	91.52	0.9634
MSPM	87.39	93.51	92.63	94.18	0.9885

To confirm the advantages of the MSPM model across various identification metrics are not accidental, this article conducts statistical significance tests on key classification accuracy indicators Acc-5, Acc-3, and AUC in Table 2. For the overall performance comparison between the models,

this paper employs the paired sample t-test. This method is suitable for comparing the mean differences of performance indicators of two models on the same test set samples. The paper also reports the non-parametric Wilcoxon signed-rank test results as a robustness verification, and the conclusions of both are consistent. The significance level is set at 0.05. All performance indicators are calculated and compared based on the average values of the results from 5 independent runs to control the randomness during the training process. As shown in Table 3, the performance differences of the MSPM model compared to all baseline models in each indicator have p-values far less than 0.05. This indicates with 95% confidence that the performance improvement of the proposed MSPM model is statistically significant.

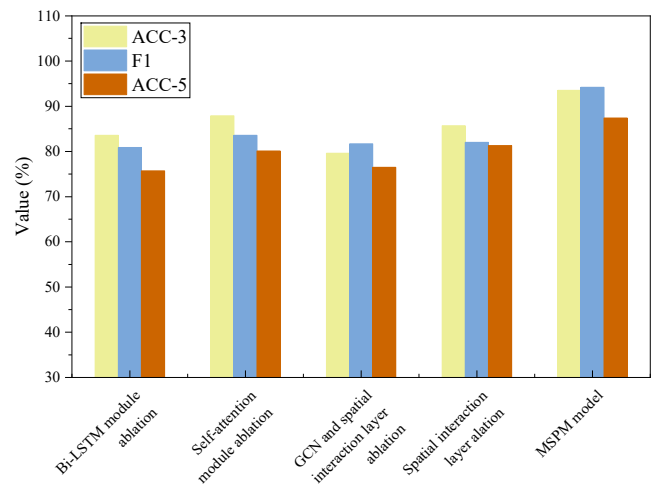
Table 3 Statistical significance test results

Comparison group (MSPM vs.)	P-value			Whether significant
	Acc-5	Acc-3	AUC	
CNN-RNN	0.035	0.018	0.014	Yes
MSTL	0.029	0.015	0.012	Yes
MSCNN	0.025	0.011	0.007	Yes
MIMN	0.019	0.009	0.005	Yes
AMTCNN	0.016	0.005	0.001	Yes

5.4 Analysis of ablation experiment results

For the goal of further exploring the contribution of different modules to the MSPM model, this section designs multiple sets of ablation experiments. Firstly, this paper assesses the impact of removing the key modules in the model (module ablation). Secondly, in order to verify the effectiveness of the proposed dynamic layer weighting strategy based on Pearson correlation coefficient, we compare it with three common layer feature fusion strategies (strategy ablation).

Figure 6 Ablation experiment results (see online version for colours)



5.4.1 Module ablation experiment

To further explore the contribution of different modules to the MSPM model, this section designs four sets of ablation experiments for various components. In each experiment, different modules are removed respectively to validate the effectiveness of each module. The evaluation metrics selected are Acc-5, Acc-3, and F1.

- 1 BiLSTM module ablation: remove the time series feature extraction module from the MSPM model, and directly input the feature vectors without temporal information into the semantic interaction layer.
- 2 Self-attention module ablation: remove the self-attention module in the MSPM model. The context features obtained by the BiLSTM module are directly fed into GCN for feature alignment and semantic interaction, bypassing the self-attention mechanism for weight matching.
- 3 GCN and spatial interaction layer ablation: remove the GCN module and spatial interaction layer from the semantic interaction layer. Directly fuse and classify the features obtained through the self-attention mechanism.
- 4 Spatial interaction layer ablation: remove the spatial interaction layer of the MSPM model, aggregate the features using GCN without performing dynamic allocation for convolutional layer weights, then directly perform fusion recognition.

As shown in Figure 6, the ablation experiment of the GCN module and spatial interaction layer also demonstrates that this module has the greatest impact on the SST-DIN model, leading to an 8.3% decrease in Acc-3 and a 9.1% decrease in F1. The self-attention module still has the smallest effect on the overall performance of the model, causing decreases of 1.5% in both Acc-3 and F1 values. Ablations of other modules also affect the model's accuracy to varying degrees. Therefore, the absence of any module results in a loss of precision for the MSPM model, with the complete MSPM model showing the best performance.

5.4.2 Strategy fusion abandonment experiment

To demonstrate the superiority of the dynamic layer weighting strategy proposed in this paper, while keeping the other parts of the model unchanged, we replaced the feature fusion strategy of the GCN part with the following three baseline methods for comparison.

- 1 Uniform Avg: uniform layer average. Directly perform arithmetic averaging on the output features of the K-layer GCN.
- 2 Attn-weight: attention-based weighting. Introduce a learnable attention network to generate scalar weights for each layer, and then perform weighted summation.
- 3 Last-layer: directly use the output of the last layer of the GCN as the graph representation.

This paper compares the above three baseline strategies with the proposed dynamic weighting strategy based on Pearson correlation (Pearson-weight). The results are shown in Table 4. Compared with uniform avg., the proposed method has increased Acc-3 by 4.97 percentage points and F1 by 7.13 percentage points. This proves that simple averaging fails to distinguish the contribution differences of different convolution layers in specific contexts, while the proposed dynamic weighting mechanism can allocate weights more precisely. Compared with Attn-weight, the proposed method has increased Acc-3 by 1.18 percentage points and F1 by 1.77 percentage points. This verifies that explicit weighting based on statistical correlation, which is learned explicitly by the attention mechanism, can produce more interpretable and stable importance measures, thereby achieving better performance. compared with last-layer, the Pearson-weight method has the most obvious advantage. This indicates that in the task of tourism intent recognition, different levels of spatiotemporal features are complementary. Relying solely on the last layer will lose valuable information, while Pearson-Weight can effectively integrate multi-level information.

Table 4 Performance comparison of different layer fusion strategies

Integration strategy	Acc-5	Acc-3	F1	AUC
Uniform Avg.	82.71	88.54	87.05	0.9368
Attn-weight	86.05	92.33	92.41	0.9812
Last-layer	85.12	91.05	91.27	0.9743
Pearson-weight	87.39	93.51	94.18	0.9885

5.5 Error analysis and discussion

To deeply understand the limitations of the MSPM model and identify the future improvement directions, this paper conducts a systematic analysis of the model's prediction error cases. The main error patterns are summarised as follows.

- 1 Confusion caused by semantic ambiguity between categories. Analysing the confusion matrix reveals that the model has a high mutual misjudgment rate in the two intents of theme tourism and leisure vacation. These two intents inherently overlap in the real user expressions, and their decision boundaries are inherently fuzzy. Future work can introduce more granular annotations or hierarchical classifiers to handle such issues.
- 2 Errors caused by missing modalities or low quality. In some cases of prediction errors, this paper finds that the multi-modal data is incomplete or of low quality. For example, a text mainly expressing an adventurous exploration intent, if the accompanying picture is blurry or the voice background is noisy, the model may rely more on the text modality. This reflects that the current

model still has room for improvement in its robustness to modal noise and absence.

- 3 The influence of imbalanced data distribution, with an unbalanced distribution of categories in the dataset. Although the overall accuracy of the model is high, the recall rate for the minority classes is relatively low. This indicates that the patterns learned by the model from sufficient data dominate the decision-making, and the learning of minority class patterns may not be sufficient. Adopting class-balanced sampling or weighted loss functions is a potential improvement strategy.
- 4 Insufficient understanding of complex spatiotemporal correlations. A few error cases involve misjudgments of long-range, complex spatiotemporal logic. For example, a user expressed preferences for both leisure and compact schedules in multiple conversations. The model may only capture local preferences but fail to integrate contradictory unified intentions. Future research needs to enhance the modelling ability for long sequences, contradictions, or variable intentions.

This in-depth error analysis not only identifies the error types but also links the classification confusion with the potential internal failure mechanisms of the model. The analysis shows that future improvements can be made in two directions. One is to enhance the robustness of the fusion module to modal conflicts or noise, such as introducing modal confidence estimation or adversarial training to reduce the weight of unreliable modalities. The other is to optimise the cross-modal alignment algorithm and explore alignment strategies with stricter semantic constraints to construct a purer, semantically consistent graph structure for reasoning.

6 Conclusions

With the vigorous development of tourism, the demand of tourists for personalised and intelligent tourism services is increasing day by day. Traditional methods for understanding and planning routes mostly rely on single-modal data or simple rules, which are unable to cope with complex and variable tourism scenarios and diverse user needs. Therefore, this paper proposes a general tourism intent understanding and route planning method based on a multi-modal spatio-temporal pre-training model. The main contributions of this method are summarised as follows.

- 1 A heterogeneous graph construction method based on timestamp alignment is proposed. The temporal synchronicity of multi-modal data is transformed into graph structure constraints, achieving precise modelling of cross-modal spatio-temporal features. This method provides a new solution to the problem of temporal multi-modal alignment.
- 2 A dynamic weighted mechanism based on Pearson correlation coefficient is designed for the convolution

layer. Through the statistical correlation analysis of feature distributions, the adaptive importance assessment of the graph convolution layer is achieved, breaking through the limitations of traditional fixed-weight fusion.

- 3 A complete spatio-temporal multi-modal fusion framework is constructed. The above innovative mechanisms are integrated organically to form a full-process innovation from feature extraction, temporal modelling, semantic interaction to spatial fusion.

Experimental results show that this research significantly improves the accuracy of intent recognition and the efficiency of route planning, providing a new technical paradigm for intelligent tourism services.

7 Limitations and future work

This study has achieved certain phased results in constructing a framework for understanding general tourism intentions and route planning. However, it is necessary to point out that there are inherent limitations in the data, scenarios, and application verification aspects. These boundary conditions provide important context for accurate assessment and subsequent improvement.

- 1 Dataset bias and representativeness limitations. The multimodal dataset relied upon for the experiment originated from a single online travel platform. The user group, content style, and covered tourism types of this dataset have sampling biases specific to the platform and cultural circle. This may lead to a decline in the model's generalisation ability when dealing with cross-platform, cross-cultural, or niche tourism scenarios. Future efforts should focus on constructing a balanced dataset from multiple sources and across cultures to alleviate this issue.
- 2 Geographical and scenario modelling limitations. The construction and verification of the route planning module are based on typical tourist cities and scenic spots within China. The transportation mode, cost structure, and attributes of the attractions reflect specific regional characteristics. When the model is directly transferred to other regions with different transportation networks, consumption levels, or tourism ecosystems, its core assumptions and parameter settings may fail, limiting its global universality.
- 3 Lack of real-time system deployment and online verification. All evaluations are currently conducted on offline historical datasets, which are retrospective simulation validations. The research has not yet undergone the test in a real online environment, including: dealing with dynamic interactions and real-time evolution of user intentions, integrating rapidly changing real-world data, and ensuring the response performance and stability of the system under

high concurrent requests. This is a key gap from the academic prototype to a reliable service.

In response to these limitations and to further advance the field to a deeper level, we propose the following three closely related and feasible future research directions.

- 1 Explore an end-to-end multimodal unified encoding and fusion framework based on Transformer. By treating multimodal data as a unified labelled sequence. And using the cross-attention mechanism to dynamically learn modality alignment and interaction, the flexibility and representational ability of the model can be enhanced.
- 2 Break through the limitations of the existing two-stage separate design and study an integrated model that can be jointly trained end-to-end. Utilise the reinforcement learning framework based on policy gradients, or build differentiable planning approximators, so that the intention understanding module can directly optimise based on the final planning utility, achieving consistent global goals.
- 3 Conduct real-world deployment and user-centred evaluation. The key step in promoting research results from the laboratory to practical applications is to build an online prototype system and conduct rigorous field user studies. Develop interactive applications for small-scale pilot deployments, and collect real user behaviour data and subjective feedback through A/B testing, interviews, and long-term satisfaction surveys. This not only provides more reliable guidance for model optimisation than offline metrics, but is also the only way to verify its practical value and discover unforeseen usage patterns.

Acknowledgements

This work is supported by the Henan Social Science Planning Special Project named: Research on the Implementation of the Integration Strategy of Culture, Tourism, Culture and Creativity in Henan Province (No. 2022ZT29) and Henan Province Soft Science Research Program Project (No. 262400410379).

Declarations

All authors declare that they have no conflicts of interest.

References

- Absalamova, G.Z., Kamalov, S. and Absalamova, D. (2025) 'Implementation and evaluation of heuristic algorithms for real-time tourist route planning', *American Journal of Business Management, Economics, and Banking*, Vol. 34, pp.35–48.
- Arif, S. and Bashir, H. (2024) 'From narratives to destinations: semantic-spatial modeling of tourism trends using geotagged reviews', *Frontiers in Computational Spatial Intelligence*, Vol. 2, No. 1, pp.43–53.
- Azarberahman, A., Tohidinia, M. and Aliakbarzadeh, H. (2025) 'Multi-objective portfolio optimization strategy using the SPEA-II algorithm', *Journal of Information and Communication Technology*, Vol. 24, No. 1, pp.52–78.
- Beed, R., Roy, A., Sarkar, S. and Bhattacharya, D. (2020) 'A hybrid multi-objective tour route optimization algorithm based on particle swarm optimization and artificial bee colony optimization', *Computational Intelligence*, Vol. 36, No. 3, pp.884–909.
- Bolotbekova, A., Hakli, H. and Beskirli, A. (2025) 'Trip route optimization based on bus transit using genetic algorithm with different crossover techniques: a case study in Konya/Türkiye', *Scientific Reports*, Vol. 15, No. 1, pp.24–31.
- Calderón-Fajardo, V., Rodríguez-Rodríguez, I. and Puig-Cabrera, M. (2025) 'From words to visuals: a transformer-based multi-modal framework for emotion-driven tourism analytics', *Information Technology and Tourism*, Vol. 27, No. 4, pp.939–979.
- Castro, F.M., Marin-Jimenez, M.J., Guil, N. and Pérez de la Blanca, N. (2020) 'Multimodal feature fusion for CNN-based gait recognition: an empirical comparison', *Neural Computing and Applications*, Vol. 32, No. 17, pp.14173–14193.
- Chen, J., Cong, J., Li, M., Sun, Y. and Zhang, J. (2025) 'T-ECBM: a deep learning-based text-image multimodal model for tourist attraction recommendation', *Scientific Reports*, Vol. 15, No. 1, pp.41–63.
- Gupta, V., Jung, K. and Yoo, S.-C. (2020) 'Exploring the power of multimodal features for predicting the popularity of social media image in a tourist destination', *Multimodal Technologies and Interaction*, Vol. 4, No. 3, pp.64–79.
- He, S. (2023) 'A novel travel route planning method based on an ant colony optimization algorithm', *Open Geosciences*, Vol. 15, No. 1, pp. 54–61.
- Hu, T. and Geng, J. (2024) 'Research on the perception of the terrain image of the tourism destination based on multimodal user-generated content data', *PeerJ Computer Science*, Vol. 10, pp.80–97.
- Huang, X., Ma, T., Jia, L., Zhang, Y., Rong, H. and Alnabhan, N. (2023) 'An effective multimodal representation and fusion method for multimodal intent recognition', *Neurocomputing*, Vol. 548, pp.12–26.
- Huang, Y., Ma, L. and Wang, Y. (2025) 'Multi-layer prompt engineering for intent recognition in travel planning assistants', *Applied Sciences*, Vol. 15, No. 21, pp.42–55.
- Kwon, D.H. and Yu, J.M. (2024) 'Real-time multi-CNN-based emotion recognition system for evaluating museum visitors' satisfaction', *ACM Journal on Computing and Cultural Heritage*, Vol. 17, No. 1, pp.1–18.
- Li, Q., Li, S., Hu, J., Zhang, S. and Hu, J. (2018) 'Tourism review sentiment classification using a bidirectional recurrent neural network with an attention mechanism and topic-enriched word vectors', *Sustainability*, Vol. 10, No. 9, pp.33–42.
- Liu, Y., Jiang, Y. and Luo, Q. (2024) 'Advancing tourism resilience and data science using multimodal data', *Journal of Policy Research in Tourism, Leisure and Events*, Vol. 4, pp.1–11.

- Marks, M., Knott, M., Kondapaneni, N., Cole, E., Defraeye, T., Perez-Cruz, F. and Perona, P. (2025) 'A closer look at benchmarking self-supervised pre-training with image classification', *International Journal of Computer Vision*, Vol. 4, pp.1–13.
- Miyazawa, K., Kyuragi, Y. and Nagai, T. (2022) 'Simple and effective multimodal learning based on pre-trained transformer models', *IEEE Access*, Vol. 10, pp.29821–29833.
- Pantano, E. and Styliadis, D. (2021) 'New technology and tourism industry innovation: evidence from audio-visual patented technologies', *Journal of Hospitality and Tourism Technology*, Vol. 12, No. 4, pp.658–671.
- Perezhohin, Y., Santos, T., Costa, V., Peres, F. and Castelli, M. (2024) 'Enhancing automatic speech recognition: effects of semantic audio filtering on models performance', *IEEE Access*, Vol. 12, pp.155136–155150.
- Qi, T., Xu, Y. and Ling, H. (2019) 'Tourism scene classification based on multi-stage transfer learning model', *Neural Computing and Applications*, Vol. 31, No. 8, pp.4341–4352.
- Sharma, M.P., Meena, U. and Sharma, G.K. (2022) 'Intelligent data analysis using optimized support vector machine based data mining approach for tourism industry', *ACM Transactions on Knowledge Discovery from Data (TKDD)*, Vol. 16, No. 5, pp.1–20.
- Si, X., Wu, X., Sheng, H., Zhu, J. and Li, Z. (2024) 'SeisCLIP: A seismology foundation model pre-trained by multimodal data for multipurpose seismic feature extraction', *IEEE Transactions on Geoscience and Remote Sensing*, Vol. 62, pp.1–13.
- Stoyanova-Doycheva, A., Ivanova, V., Glushkova, T., Stoyanov, S. and Radeva, I. (2020) 'Dynamic generation of cultural routes in a tourist guide', *International Journal of Computing*, Vol. 19, No. 1, pp.39–48.
- Wang, X., Chen, G., Qian, G., Gao, P., Wei, X.-Y., Wang, Y., Tian, Y. and Gao, W. (2023) 'Large-scale multi-modal pre-trained models: a comprehensive survey', *Machine Intelligence Research*, Vol. 20, No. 4, pp.447–482.
- Wu, X., Guan, H., Han, Y. and Ma, J. (2017) 'A tour route planning model for tourism experience utility maximization', *Advances in Mechanical Engineering*, Vol. 9, No. 10, pp.16–29.
- Yan, J., Yue, J., Zhang, J. and Qin, P. (2023) 'Research on spatio-temporal characteristics of tourists' landscape perception and emotional experience by using photo data mining', *International Journal of Environmental Research and Public Health*, Vol. 20, No. 5, pp.38–53.
- Zhang, Y. and Sun, Y. (2024) 'Smart travel planning system based on kernel density estimation and similarity metric clustering algorithm', *Evolutionary Intelligence*, Vol. 17, No. 5, pp.4227–4238.
- Zhou, C. and Huang, M. (2024) 'The persuasive effects of voice characteristics embedded in paid tour guide audio on tourist purchase decisions based on deep learning', *Journal of Hospitality and Tourism Management*, Vol. 60, pp.313–321.