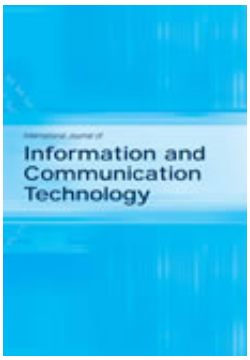
INDERSCIENCE PUBLISHERS
Linking academia, business and industry through research



International Journal of Information and Communication Technology

ISSN online: 1741-8070 - ISSN print: 1466-6642
https://www.inderscience.com/ijict

A CNN-ViT fusion model for predicting tourist behaviour and consumption intentions

Chiawei Liu

DOI: 10.1504/IJICT.2026.10075839

Article History:

Received:	28 October 2025
Last revised:	01 December 2025
Accepted:	02 December 2025
Published online:	02 February 2026

A CNN-ViT fusion model for predicting tourist behaviour and consumption intentions

Chiawei Liu

School of Geographical Sciences and Tourism,
Jiaying University,
Meizhou, 514015, China
Email: liuinjyu@163.com

Abstract: Facing the problem that current methods have in inferring tourist behaviour and consumption intentions due to the difficulty in deeply mining textual semantic information, this paper optimises convolutional neural networks-based vision transformer algorithm first, and then proposes improved inference tourist behaviour and consumption intentions based on Visual Transformer. In tourist behaviour detection branch, text-aware module is introduced to improve the extraction of tourist image features and enhance the expressive power of textual visual features. In consumption intention inference branch, parallel transformer decoding is performed at both visual and linguistic levels, and semantic information is mined and integrated by positional encoding to realise accurate inference of consumption intentions. The experimental results show that the accuracy of visitor behaviour detection is 96.8%, and the accuracy of consumption intention inference is 94.2%. Compared with the baseline model, the model has high efficiency and is superior.

Keywords: convolutional neural network; CNN; vision transformer; feature extraction; tourist behaviour detection; consumption intent inference.

Reference to this paper should be made as follows: Liu, C. (2026) 'A CNN-ViT fusion model for predicting tourist behaviour and consumption intentions', *Int. J. Information and Communication Technology*, Vol. 27, No. 2, pp.83–99.

Biographical notes: Chiawei Liu is an Associate Professor in the School of Geographical Sciences and Tourism at Jiaying University, China. He received a PhD from Peking University, Beijing, China, in 2021. He published three SSCI index papers. His research interests include tourism consumer behaviour, tourism survey and statistics.

1 Introduction

With the rapid growth of global tourism, it is very important to have a good understanding of tourist behaviour and correctly identify the tourist's consumption intent. Most of the previous researches on tourist behaviour and consumption intent inference are based on subjective methods, such as surveys and interviews (Choe and Kim, 2018). Although these methods can acquire rich users' psychological motivations, they still have some limitations, such as small sample size and poor real-time performance (An et al.,

2022). A large amount of tourism data is constantly generated in the digital age, which offers new opportunities to analyse tourists. Traditional tourist behaviour analysis still mostly depends on machine learning models based on manually designed features (Zhang et al., 2025). These methods have played an important role in handling small-scale structured data. However, when facing high-dimensional unstructured visual and space-time data, these methods are limited in their feature representation ability and are strongly dependent on human expertise. It is difficult to model the complex non-linear patterns in the data (Rezapouraghdam et al., 2023). With the emergence of deep learning technology, which has powerful ability of end-to-end feature learning and hierarchical abstraction, it has brought revolutionary breakthroughs to this field. However, consumer intent is a higher-level and abstract semantic concept. It is not determined by a single isolated behaviour, but exists in a series of temporally combined behaviours and persistent attention to certain goals (Wu et al., 2025). How to use deep learning algorithms to infer tourist behaviour and consumption intent is a highly challenging research problem.

Inferring tourist behaviour and consumption intent is very important for the tourism industry to provide personalised travel service. Yu et al. (2025) explored the influencing factors of tourist consumption intent, and found that product usage times affect consumers' willingness to consume. However, this method may overlook other potential consumption influencing factors. Senbeto and Hon (2020) proposed a deep adaptive evolutionary ensemble model for predicting tourist behaviour and consumption intent, which introduces model diversity into the cascade layer to adapt to complex purchasing behaviour patterns. However, the complexity of the model may lead to increased computational costs. Kang et al. (2022) optimised the XGBoost classifier through feature selection and oversampling methods, although this approach performs well in improving prediction performance, it may result in the loss of key information. Sharma et al. (2022) proposed tourist behaviour and consumption intent inference method based on improved support vector machines, but the inference error of this method is relatively large. Gregoriades et al. (2023) analysed tourists' social platform tweets using a random forest model to identify potential consumption intentions, but tweet data may contain noise, and tourist expressions may be ambiguous, which can affect the accuracy of the analysis.

Machine learning-based methods for inferring tourist behaviour and consumption intent can extract valuable information from large amounts of data, but tourist behaviour involves a large number of random factors, which are difficult for models to capture in real-time. Additionally, machine learning algorithms rely on a large amount of labelled data, and the high cost and subjectivity of manual labelling lead to low inference accuracy of the models. The research based on deep learning does not need to design features, because the model can directly extract the global and local features from raw data and mine the hidden correlations automatically. In addition, the deep learning algorithm has the ability to model temporal and dynamic behaviours, which greatly improve the inference accuracy of tourist behaviour and consumption intent. Deep learning algorithms construct understanding through hierarchical abstraction, grasp relationships via distributed representations, capture dynamics through temporal modelling, focus on key elements using attention mechanisms, and achieve deep integration through end-to-end learning. Ultimately, they transform a vaguely defined business problem into a computational problem that can be efficiently solved by data-driven models. This is precisely the fundamental reason behind their exceptional inference performance. Bai and Han (2020) designed a multi-scale CNN to extract the

global and local features from tourist tweets and used an attention mechanism to amplify the key features to tourist behaviour recognition, but the recognition accuracy was only 76.9%. Mou and Wang (2025) designed an adaptive sparse attention network, used an adaptive sparse channel attention module to capture the detail change caused by external disturbance to enhance the robustness, and did a good job in inferring tourist behaviour and consumption intent. Rong et al. (2024) improved the temporal convolutional network (TCNs) by using a self-attention mechanism to further increase the weight of key features to eliminate the impact of information on recognition results. Methods of inferring tourist behaviour and consumption intent based on ensemble models are to combine and adjust multiple recognition methods, use the advantages of multiple models to maximise the advantages of each method, and improve the accuracy of prediction (Ye and Huang, 2022). Liu and Hu (2025) proposed a method of inferring tourist consumption intent based on long short-term memory (LSTM) and visual transformer, achieving more accurate inference results compared with single deep learning method. Shao and Kim (2020) proposed a tourist behaviour and consumption intent prediction method based on CNN and LSTM, aiming to solve the poor prediction results of redundant data, so as to improve the generalisation ability of model. Si (2025) proposed hybrid model, integrating a gated recurrent unit (GRU) and a vision transformer (ViT), first detects tourist behaviour with the GRU and then infers consumption intent with the ViT, resulting in excellent inference accuracy.

Based on a detailed analysis of the above research work, it can be found that the current research ignores the spatiotemporal complexity of tourist behaviour and the implicit semantic association between behaviour and consumption decision, which leads to the low accuracy of inferring tourist behaviour and consumption intent. Therefore, this paper proposes a model for inferring tourist behaviour and consumption intent based on CNN and ViT algorithm. The model first optimises the visual transformer algorithm by introducing the hierarchical structure of convolutional neural network (CNN). And then the improved visual transformer based model for inferring tourist behaviour and consumption intent is designed. The model is composed of tourist behaviour detection branch and consumption intent inference branch. In the detection branch, the text-aware module is especially introduced. The text-aware module enhances the extraction ability of tourist image features and reduces the interference of background noise, which improves the text-visual features. In the inference branch, parallel transformer decoding is performed from both the visual and linguistic levels, and semantic information is fully mined and integrated through positional encoding, thus achieving precise inference of consumption intent. Experimental results show that the proposed model's accuracy in inferring tourist behaviour and consumption intent is at least 2.7% and 3.8% higher than that of the baseline model, significantly improving the accuracy of inferring tourist behaviour and consumption intent.

2 Relevant theory

2.1 Convolutional neural network

CNN is a neural network specifically designed for processing grid-structured data, particularly suitable for handling images and spatially related data. Its core feature is the ability to automatically learn local features, possessing strong feature extraction

capabilities (Zhang et al., 2019). CNN can extract local features from data through convolutional layers, reduce feature dimensions through pooling layers, and finally complete classification tasks through fully connected layers. The core components of CNN include convolutional layers, pooling layers, and fully connected layers (Zhao and Zhang, 2024).

The convolutional layer extracts the local feature by sliding the convolutional kernel over the input data. Each convolutional kernel performs dot product operation on a local region of the input and outputs a feature map. The pooling layer downsamples the feature map, reducing the data dimensionality while also lowering computational complexity and the risk of model overfitting. Fully connected layers map the input data through weights and activation functions. Their main function is to map the features extracted by the preceding convolutional and pooling layers to the output space, thereby achieving high-level feature integration and decision-making (Zhang et al., 2024). The advantage of CNNs lies in their ability to process data in parallel, making them highly efficient. However, since convolutional kernels can only capture local features, CNNs are also challenged in efficiently performing global modelling. In practical applications, the CNN architecture is adjusted according to the specific needs of the task. For larger-scale tasks, the model must include more layers and more complex connections.

2.2 Vision transformer

ViT is a pioneering model that migrates the transformer architecture from natural language processing (NLP) to the field of computer vision, breaking the long-term dominance of CNNs in visual tasks. By directly processing image data through the self-attention mechanism, it demonstrates powerful feature extraction and generalisation capabilities. Unlike traditional CNNs, ViT does not rely on local convolutional operations to extract features but instead uses the global modelling capability of transformers to process images directly. ViT first divides the input image X into several non-overlapping image patches of a fixed size of $P \times P$, resulting in a total of $N = HW/P^2$ patches. Each patch is flattened into x_p and mapped to a fixed dimension D through a linear projection, obtaining the image patch embeddings as shown in equation (1).

$$z_0^p = x_p W_E + E_p \tag{1}$$

where W_E is the linear mapping weight, and E_p is the learnable position encoding used to retain positional information. On this basis, the ViT model adds a special classification (CLS) token at the beginning of the sequence, denoted as $z_0^{[CLS]}$, for the final classification task. The entire input sequence is shown in equation (2), where E_{pos} is the position encoding of the entire sequence.

$$Z_0 = \left[z_0^{[CLS]}; z_0^1; z_0^2; ...; z_0^N \right] + E_{pos} \tag{2}$$

ViT uses a standard transformer encoder architecture in the encoding stage, with its core modules including layer normalisation (LN), multi-head self-attention mechanism (MSA), and multilayer perceptron layers, ultimately outputting prediction results through the classification head (Li et al., 2022). When there is a sufficient amount of data for training, the ViT model achieves performance on par with or even surpasses

contemporary state-of-the-art models on multiple image recognition datasets (Yao et al., 2023). In general, the application of ViT requires pre-training on a large amount of data, and fine-tuning on targeted small datasets during deployment to adapt to downstream tasks.

Traditional ViT models directly segment images into image patches and treat them as sequences fed into the transformer, lacking the inductive biases inherent in CNNs such as local connections and weight sharing. Inductive biases help models better learn intrinsic patterns in data with limited samples. Consequently, ViT requires substantial training data to achieve good performance; when data is scarce, it tends to overfit and exhibits poor generalisation. The convolutional kernels in CNNs feature local connectivity, with each kernel focusing only on a specific local region of the input data. This enables effective extraction of local image features such as edges and textures. Incorporating CNNs into ViT allows the model to leverage convolutional operations to enhance its ability to capture local information, thereby improving its understanding of image details. After integrating CNN into ViT, images undergo preliminary feature extraction via CNN, converting high-resolution images into low-resolution feature maps. These feature maps are then segmented into smaller sequences for input to the transformer. This approach reduces the number of image patches, thereby lowering computational complexity for the self-attention mechanism and accelerating both model training and inference speeds.

Applying ViT to tourist behaviour image recognition derives its core value from shifting from the 'local feature-driven' CNN paradigm to the globally context-driven transformer paradigm. This shift enables the model to understand behaviours by observing the entire scene and the intricate relationships between its elements – much like humans do – rather than relying solely on local appearance patterns. For behaviour recognition tasks in complex, open, and dynamic tourism environments, this represents a qualitative leap, significantly enhancing the model's accuracy, robustness, and intelligence.

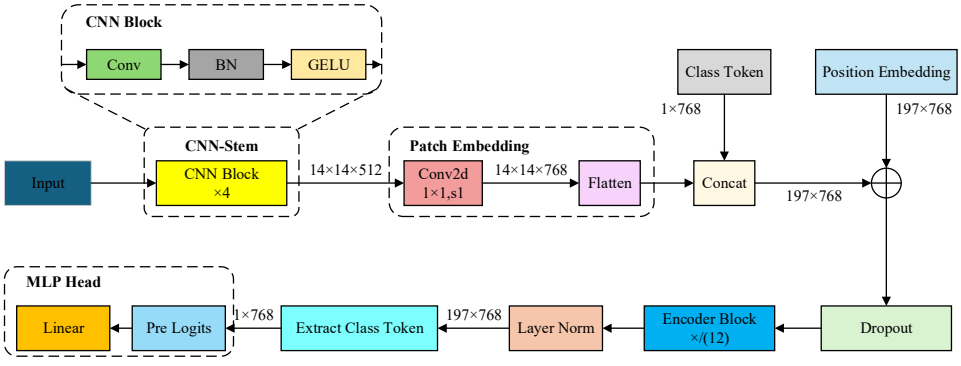
3 ViT algorithm optimisation based on CNN

The ViT algorithm breaks the traditional neural network approach to processing sequential data by introducing the self-attention mechanism to compute and model the relationships between sequence information. ViT divides the original image into equal-sized patches and projects them into tokens, allowing images to be processed similarly to sentences in NLP. In most cases, the classification target in the original image is located at the centre, and directly partitioning the entire image may result in many patches carrying information unrelated to the core of the original image. Even after multiple transformer encoder modules, while information interaction is possible, this also increases the difficulty of model training. In contrast, CNNs have a well-structured hierarchical structure that facilitates smooth feature extraction. Therefore, to address the shortcomings of the ViT algorithm and leverage the advantages of CNNs, the ViT algorithm is improved, and the resulting model is named CNViT.

The structure of the CNViT model is shown in Figure 1. The original image is fed into the CNN-Stem module, which consists of four consecutive CNN blocks. Each block internally comprises a convolutional layer, batch normalisation (BN), and a GELU activation function stacked in sequence. For the convolutional layer in each CNN Block, a 3 × 3 kernel size is used with a stride of 2, and appropriate padding of 1 is set to

maintain the coherence and completeness of the feature map size. This setup not only can capture the local spatial features but also can gradually reduce the spatial dimension of feature map when down-sampling, finally converting the original image into one feature map of size 14 × 14.

Figure 1 CNViT model structure (see online version for colours)



In the application stage of ViT module, the feature map output from CNN-Stem module will firstly go through patch embedding layer. Firstly, split the image into several fixed-size patches, and use a convolutional layer with smaller number of channels between patches to adjust, then form a new feature map. After that, flatten the spatial dimension of feature map and add one learnable class token on top of it. Class token is concatenated with the flattened feature vector. To preserve the position of every patch during the combination process, position embedding is used. It is done by adding position embedding vector to each element of patch feature vector. Main structure of the model is built by the following 3 parts above. Afterwards, the feature vectors integrated with positional information and class token pass through a dropout layer to reduce the risk of overfitting, and then undergo a series of encoder layers with self-attention mechanisms to enhance the interaction and learning among features. Meanwhile, LayerNorm layers are used to normalise the features, improving model stability and generalisation. In this process, the focus is on the output of the class token after layer-by-layer encoding, as it carries high-level abstract information of the entire sequence. Finally, the output of the class token is further transformed and classified through the MLP head layer. This series of complex operations is aimed at fully leveraging the advantages of CNN and ViT to improve the accuracy of classification tasks.

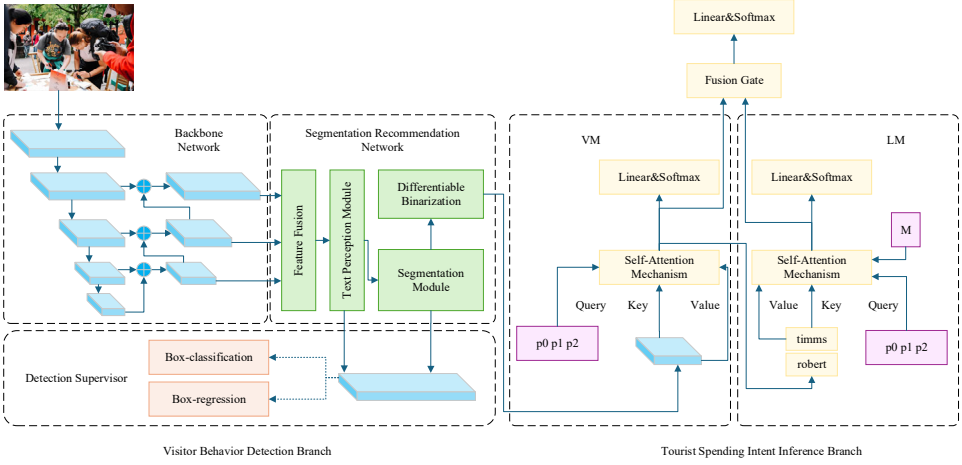
4 Inference of tourist behaviour and consumption intentions based on the CNViT algorithm

4.1 Structure of the model for inferring tourist behaviour and consumption intentions

Data for inferring tourist behaviour and purchase intent is primarily found on social media platforms, online travel agencies, and travel forums. This data not only includes single-modal visual or textual semantic information but also multimodal behavioural data. Existing research on inferring tourist behaviour from social media posts has been

affected by the lack of fixed text position, diverse font and colour variations, leading to inaccurate inference results. To address these issues, this paper designs a tourist behaviour and purchase intent inference model based on the CNViT algorithm, as shown in Figure 2. The model comprises three key components that allow for the accurate inference of both tourist behaviour and purchase intent.

Figure 2 The suggested tourist behaviour and purchase intent inference model (see online version for colours)



1 Backbone network: it uses CNN as its backbone to extract rich multi-level visual features from tourist behaviour image. Not only these basic visual features include colour, texture, shape, but also these visual features contain rich semantic information. They provide very important clues for tourist purchase intent inference. Main network will extract features step by step through multiple layers convolution, deeply mining the image to get the deep features. They are fine-grained features that can deeply distinguish text from background and text from non-text objects. It makes the whole subsequent recognition possible.

2 Tourist behaviour detection branch: it is used to accurately locate the text region in the image of tourist behaviour. It contains two main sub-modules: segmentation network and detection supervisor. The segmentation network uses advanced deep learning segmentation method to divide the image finely, and accurately find the candidate area that may contain the text. The detection supervisor will play a very important role in the training process to provide supervisory signal for the detection.

3 Purchase intent inference branch: it uses the visual module(VM) and language module(LM) of CNViT algorithm and add a fusion gate to deeply understand and recognise the text content. VM module refines and improves the initial text area output from detection branch. LM module uses self-attention method to extract contextual dependencies of text and accurately segment the semantic information of text. Finally, the fusion gate deeply combines the feature from VM and LM module and extract the most valuable feature representation. It provides the final tourist purchase intent inference with comprehensive and complementary feature information.

4.2 Backbone network

As shown in Figure 1, the main network is located in the centre of our model. Its task is to extract rich bottom-up multi-level visual features from input tourist images. These visual features can represent fine-grained texture, shape and semantic information of input images, and provide basis for next purchase intent inference. In this paper, we select feature pyramid network net (FPNNet) (Liu et al., 2023) as main network to extract key visual features from input tourist images. In the process of down-sampling, we can get spatial attention module by utilising spatial relationship of intermediate text feature information. The main role of this module is to capture important information in the two-dimensional space for tourist behaviour determination. The backbone network automatically identifies key patterns from visitor behaviour data through a hierarchical feature extraction mechanism, while review text features provide subjective feedback and contextual information. The correlation between the two is achieved through multimodal fusion, dynamic updating, and noise filtering, significantly enhancing the model's predictive capability regarding visitor consumption intent. This integrated approach not only strengthens the model's robustness but also provides the tourism industry with a more precise basis for personalised services. In each convolution, a feature information matrix I is generated, and the sigmoid function is applied to it for activation, as shown in equation (3).

$$W_s(I) = \sigma f^{3\times3} Pool(I) \tag{3}$$

where $f^{3\times3}$ is the convolution operation, with a 3 × 3 convolutional kernel. During the up-sampling process, we use the uppool pooling method to extract features and generate a channel attention module to approximate the location features. Then, the features are adjusted through a shared MLP, as shown in equation (4).

$$W_c(I') = \sigma MLP\left(unpool(I)\right) = \sigma W_1 W_0 I' \tag{4}$$

where σ is the sigmoid activation function, and W_0 and W_1 are the weights of the MLP. Finally, during the feature fusion stage, the weights of the channel attention and spatial attention are combined into an attention model to enhance the feature representation capability. This process is represented as follows.

$$I = \left(W_s(I) + 1\right) \odot I \tag{5}$$

$$I'' = \left(W_c(I') + 1\right) \odot I' \tag{6}$$

where $\odot$ is the element-wise multiplication of the corresponding matrix elements. Since the sigmoid function is used to activate the elements of the attention channel, the attention module can better enhance useful image information and suppress useless information by limiting the values of the attention channel elements. Specifically, by restricting each element's value to the range [0, 1], this effect can be achieved.

4.3 Visitor behaviour detection branch

Detection branch applies convolutional module of UNet (Trebing et al., 2021) to integrate and process multi-scale visual features extracted from backbone network. The purpose of doing this is to enhance the ability of model to detect text of different sizes and

proportions, and then improve its scale robustness. Specifically, for the three kinds of feature maps that are extracted from FPNNet backbone network and have different level and receptive field, we apply 3×3 convolutional layer on each of them. Then, through upsampling operations, these features are restored to the same size, achieving scale normalisation. After the above processing, the features at each scale are concatenated together to form a multi-scale fusion feature. The fused feature can be represented as follows.

$$F = UNet(R_v) \tag{7}$$

To further improve the model's accuracy in detecting tourist behaviour, a text-aware module is designed and introduced after the visual feature fusion module of the backbone network. This module is specifically optimised for text information. The text-aware module aims to enhance the model's understanding and analysis of image semantics by effectively perceiving and selectively attending to visual features, and ensures deep and optimised feature fusion across layers through residual connections. The fused feature, after passing through the text-aware module, becomes the text-aware feature, denoted as $F_{tam} = T(F)$.

After applying the text-aware module to extract F_{tam}, this paper further refines the text feature representation. Based on this, we can define the feature map after image segmentation processing as $S = S(F_{tam})$, where the last layer of the segmentation module is a sigmoid layer, meaning the range of S is [0, 1].

Considering that text in tourist behaviour images is often blurry, making it difficult to distinguish between text and non-text regions, this paper proposes a separable binarisation formula named attention-DBNet, which incorporates threshold selection during the binarisation process into the model training, allowing the model to adaptively select the threshold. Binarising the segmentation feature map S, the binarised segmentation result is as follows, where $B_{i,j}$ is the binarised result at position (i, j), x is the value at position (i, j) of the segmentation map, and k is the magnification factor.

$$B_{i,j} = \frac{1}{1 + e^{\frac{e^{-kx} - e^{kx}}{2}}} \tag{8}$$

This paper uses the cross-entropy loss function to explain why the proposed attention-DBNet binarisation formula performs better on target images. Using equation (8) as the original function, the positive label sample loss function l_+ is shown in equation (9), and the negative label sample loss function l_- is shown in formula (10).

$$l_+ = -\ln \frac{1}{1 + e^{\frac{e^{-kx} - e^{kx}}{2}}} \tag{9}$$

$$l_- = -\ln \left(1 - \frac{1}{1 + e^{\frac{e^{-kx} - e^{k\alpha}}{2}}} \right) \tag{10}$$

Inspired by Fang et al. (2019), in the model training phase of this study, the fast R-CNN architecture is used as a detection supervisor for the tourist behaviour detection module in the training process. In this way, Fast R-CNN takes on the role of accurately detecting

performance supervision for the segmentation recommendation network, thereby enhancing the overall system's detection capability.

4.4 Branch for inferring tourist consumption intentions

In our proposed tourist behaviour and consumption intention inference model, the whole process is modelled as directly continuous from image visual features to consumption intention text semantics. While dealing with task of inferring and recognising tourists' consumption intentions, the position information of character sequence is of great importance. On one hand, it represents the order of appearance of characters; on the other hand, it also contains a lot of semantic clues. In order to make model fully use and capture the influence of position information, integatine-dimensional positional encoding into sequence decoding process based on CNViT is an effective way to improve accuracy of recognition. This paper uses the sine and cosine functions to encode positional information, which can effectively encode the position of characters in the sequence in a continuous and periodic manner, thereby enabling the model to distinguish characters at different positions, even if they have the same character content, different representations can be assigned due to positional differences. Positional encoding can be represented as follows, where *pos* represents the position index of a character in the text sequence, with its odd-dimensional index in the position vector being $2i + 1$ and its even-dimensional index being $2i$, dim represents the dimension of the position vector.

$$PE_{(pos,2i)} = sin\left(\frac{pos}{10,000^{\frac{2i}{dim}}}\right) \tag{11}$$

$$PE_{(pos,2i+1)} = cos\left(\frac{pos}{10000^{\frac{2i}{dim}}}\right) \tag{12}$$

In the design of the VM module in the tourist consumption intention inference branch, the positional encoding results obtained from formulas (11) and (12) are first used as the query input for CNViT. These positional encodings reflect the position information of characters in the text sequence. Meanwhile, the text visual features F_{tam} extracted by the text perception module are used as the key and value for CNViT. Therefore, CNViT can perform decoding operations based on the correlation between positional query information and text visual features, thus extracting high-level semantic features corresponding to each character. Finally, the text character features decoded by the VM module can be represented as follows.

$$g_v = softmax\left(\frac{QK^T}{\sqrt{C}}\right)V \tag{13}$$

where Q, K, and V are the query, key, and value, respectively, and K^T is the transpose of K. To fully comprehend and process the complex interaction information between text visual features and characters, we apply transformer module with 8 parallel attention heads. Each attention head independently capture and learn the correlation in input sequence from its own view, so as to achieve multi-dimensional understanding and parsing of characters' multi-faceted dependencies. Then, the decoded text character features g_v are processed through a linear transformation and SoftMax activation function

to generate the final recognition probability distribution, where P_v is the recognition result of the visual module, and $T_v = W_v(g_v)$, W_v are trainable hyperparameters.

$$P_v = softmax(T_v) \tag{14}$$

The LM module is very important in tourist consumption intention inference task. Its basic role is to further deeply mine and utilise the initial results obtained from visual recognition process. Inspired by the view in Koo et al. (2016), we regard tourist consumption intention inference process as a complex bidirectional understanding and fill-in-blank process, just like the bidirectional masked language model task of predicting obscured words. Through this kind of structural design, we can make use of advantages of positional encoding to extract and integrate semantic content that is most related to context within the already extracted and encoded visual features that have been extracted and encoded, and then achieve accurate inference of tourist consumption intentions.

The self-attention module in LM module applies a specially customised mask matrix M. The design of masking strategy is to simulate the operational situation of bidirectional fill-in-the-blank. In the process of encoding, make sure that the representation of current character is not depended on direct information of it, but constructed entirely by clues from context information of surrounding characters. The encoding process is represented by equation (15), where Q is the positional encoding of the character, and K and V are the key and value, respectively. The text recognition result obtained after processing by the LM module can be expressed by equation (16), where P_L is the recognition output of the language module, and $F_L = W_l(g_L)$, W_l are trainable hyperparameters.

$$g_L = softmax\left(\frac{QK^T}{\sqrt{C}} + M\right)V \tag{15}$$

$$P_L = softmax(F_L) \tag{16}$$

To effectively integrate visual features with semantic features to enhance the performance of tourist consumption intention inference, a fusion gate mechanism is set up in the model. First, the g_v encoded by the VM module and the semantic features g_L extracted by the LM module are input into the fusion gate for integration, as shown in equation (17).

$$\begin{cases} W_f = \sigma\left([g_v, g_L]W_g\right) \\ F_g = W_f g_v + (1 - W_f) g_L \end{cases} \tag{17}$$

where σ is the sigmoid activation function, used to generate a weight value between 0 and 1, which controls the relative contributions of g_v and g_L in the fusion process. $[g_v, g_L]$ denotes the concatenation of the two along a dimension to form a new feature vector, and W_g and W_f represent trainable hyperparameters and fusion weights, respectively. The prediction result of tourist consumption intention inference can be expressed as equation (18).

$$P_F = softmax(F_f) \tag{18}$$

where P_F is the predicted probability distribution output by the tourist consumption intention inference branch, and F_f is the fused feature after linear transformation,

specifically $F_f = W_r(F_g)$, and W_r are trainable hyperparameters. Through this series of operations, the model can flexibly adjust the weights of visual and semantic information based on actual conditions, thus achieving more accurate and comprehensive tourist consumption intention inference.

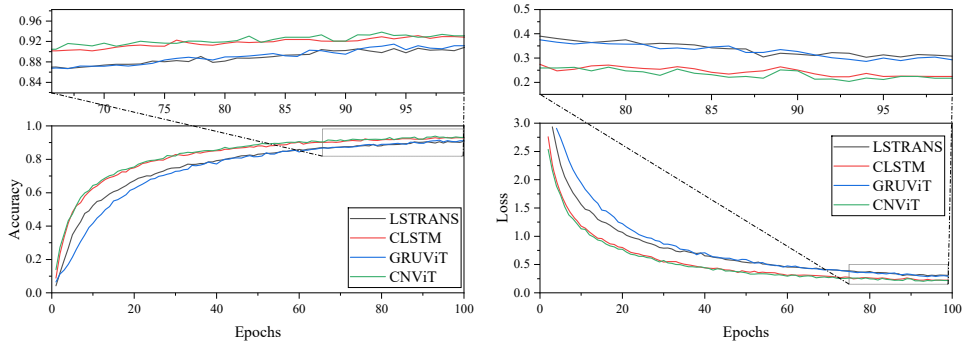
5 Experimental results and performance analysis

5.1 Analysis of tourist behaviour and inference results on their consumption intentions

This paper uses the tourist behaviour and consumption data collected in reference (Li and Cao, 2022) as the experimental dataset. This dataset includes various information on tourist behaviour from 20563 records on a travel website, such as UserID, the average number of times users browse travel-related pages annually, user reviews of travel destinations, and hotel consumption data, etc. The dataset is divided into training, testing, and validation sets at a ratio of 6:2:2. The AdamW algorithm with an initial learning rate of 0.0001 is chosen for gradient training, using a learning rate decay strategy where the learning rate decays to half of its original value every 50 epochs, with a maximum of 200 training epochs. The experiment uses an NVIDIA GeForce RTX 4060 Laptop GPU processor, and the program runs in the PyCharm software under the PyTorch 2.0.1 framework, with Python3 as the programming language.9 is the programming language.

The tourist behaviour prediction results of the proposed model CNViT are shown in Figure 3, where the x-axis represents the true labels of the model, the y-axis represents the predicted labels, and the value in each cell represents the number of samples of a certain category that the model predicts into another category. From Figure 3, it can be seen that the model performs relatively accurately in most categories, especially in the 'walk' and 'stop' categories, where 950 and 954 samples, respectively, were correctly classified, indicating very high classification accuracy in these two categories. However, the model has some confusion in certain categories, for example, 119 samples between the 'sit down' and 'rest' categories were misclassified, indicating that the model has some difficulty in distinguishing between 'sit down' and 'rest'.

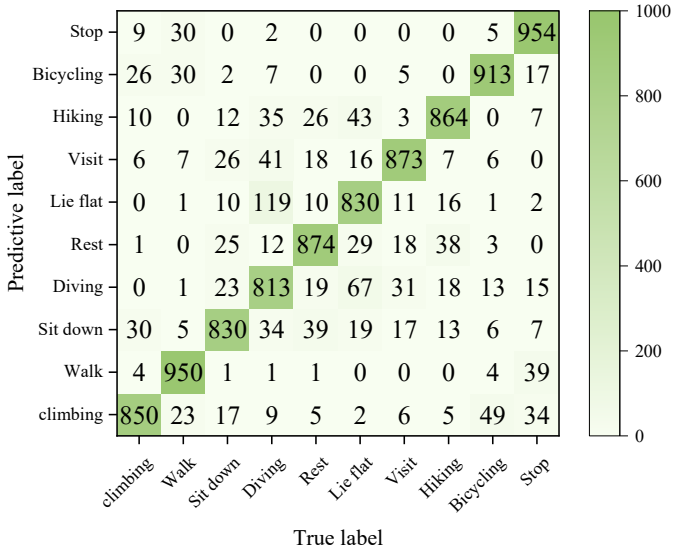
Figure 3 The tourist behaviour prediction results of the proposed model CNViT, (a) accuracy (b) loss (see online version for colours)



(a) (b)

For ease of analysis, LSTRANS (Liu and Hu, 2025), CLSTM (Shao and Kim, 2020), and GRUViT (Si, 2025) were selected as benchmark models. The inference results of tourist consumption intent for different models are shown in Figure 4. The CNViT model demonstrated the fastest convergence speed during training, achieving the highest accuracy in inferring tourist consumption intent and the lowest loss value. The accuracy of GRUViT was slightly lower than CNViT, while the loss value was slightly higher. In comparison, the LSTRANS model had the slowest convergence rate, followed by the CLSTM model.

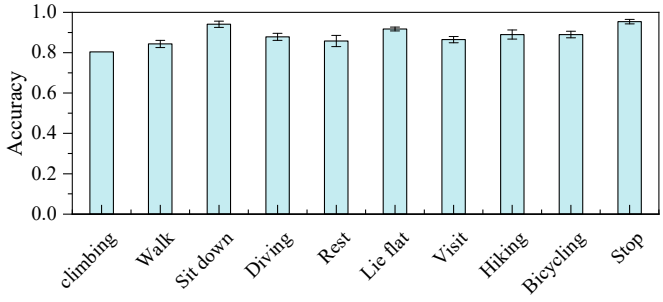
Figure 4 The inference results of tourist consumption intent for different models (see online version for colours)



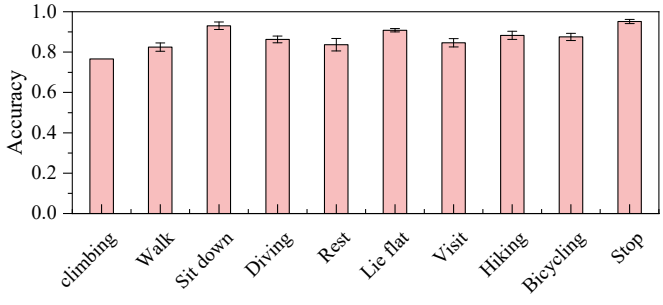
5.2 Performance analysis of different methods

The prediction accuracy of CNViT and GRUViT for various tourist behaviours is shown in Figure 5. The average prediction accuracy of CNViT for various tourist behaviours reached 96.17%, while the average prediction accuracy of GRUViT for various tourist behaviours was 95.03%. The average prediction accuracy of CNViT was 1.14% higher than that of GRUViT. Although GRUViT uses a hybrid model of GRU and ViT for predicting tourist behaviour, the GRU mitigates gradient vanishing in long sequences through gating mechanisms, but still struggles to capture ultra-long dependencies in tourists' continuous behaviours spanning several hours (such as travel routes and consumption decisions). Additionally, when GRUViT performs global self-attention calculations on tourist behaviour images, it may focus on irrelevant regions, increasing computational costs and reducing feature effectiveness. CNViT uses Unet to achieve the detection of tourist behaviour and designs a ViT segmentation network to finely divide images, accurately detecting tourist behaviour and improving the prediction accuracy of tourist behaviour.

Figure 5 The prediction accuracy of (a) CNViT and (b) GRUViT for various tourist behaviours (see online version for colours)



(a)



(b)

This paper further analyses the performance of LSTRANS, CLSTM, GRUViT, and CNViT in inferring tourist behaviour and consumption intentions using evaluation metrics such as accuracy (A), recall (R), F1, and AUC, as shown in Table 1. The accuracy of CNViT in inferring tourist behaviour and consumption intentions is 96.8% and 94.2%, respectively, which is an improvement of 9.3% and 10.3% over LSTRANS, 6.2% and 5.5% over CLSTM, and 2.7% and 3.8% over GRUViT. Comparing the harmonic mean of precision and recall, F1, CNViT's F1 scores for inferring tourist behaviour and consumption intentions are 95.9% and 96.3%, respectively, which is an improvement of 10.2% and 12.4% over LSTRANS, 7.5% and 7.2% over CLSTM, and 2.8% and 3.5% over GRUViT. Comparing the area under the ROC curve (AUC), CNViT's AUC scores for inferring tourist behaviour and consumption intentions are 0.983 and 0.979, respectively, which is at least an improvement of 0.82% and 1.87% over the other three models. LSTRANS mitigates the vanishing gradient problem of traditional RNNs through gating mechanisms, but may still fail to capture long-term dependencies due to gradient decay when processing long-term tourist behaviour trajectories. Additionally, ViT captures global information through the self-attention mechanism but may lack sufficient extraction of local details in tourist consumption scenarios, potentially affecting the judgment of consumption intentions. CLSTM typically simple concatenates the spatial features extracted by CNN and the temporal features captured by LSTM through a fully connected layer, lacking a deep feature interaction mechanism. Additionally, it may ignore the correlation between spatial and temporal features, leading to inference errors. GRUViT, which infers tourist behaviour and consumption intentions

based on GRU and ViT, but this method ignores the visual features of tourist text reviews, resulting in low inference accuracy. CNViT combines the local features extracted by CNN with the global features captured by ViT, allowing for more accurate inference of consumption intentions.

Table 1 Comparison of the inference performance for various methods

Method	*Tourist behaviour prediction*				*Consumption intention inference*			
	A	*R*	*F1*	*AUC*	*A*	*R*	*F1*	*AUC*
LSTRANS	0.875	0.842	0.857	0.902	0.839	0.846	0.839	0.898
CLSTM	0.906	0.886	0.884	0.948	0.887	0.871	0.891	0.947
GRUViT	0.941	0.905	0.931	0.975	0.904	0.918	0.928	0.961
CNViT	0.968	0.937	0.959	0.983	0.942	0.935	0.963	0.979

6 Conclusions

With the rapid development of tourism industry, large scale of tourist images and behaviour data open up new opportunities to accurately understand tourist's real preferences and consumption intentions. However, traditional methods ignore the spatiotemporal complexity of tourist behaviour and implicit semantic associations of tourist's consumption intentions, which cause the low inference accuracy. In this paper, we first optimise the ViT algorithm by integrating with hierarchical structure of CNNs. Then, we design tourist behaviour and consumption intention inference model based on optimised ViT algorithm. The two branches of tourist behaviour detection and consumption intention inference make the whole model accurately extract deep semantic information from tourist's images and efficiently decode and understand tourist's review text, which make our model get excellent performance in tourist consumption intention inference. Specifically, the consumption intention inference branch adopts the VM and LM modules as well as fusion gate component. The whole model fully extract visual features and further explore semantic information of text. Through positional encoding, visual module and language module can recognition operation together, which make the model get the advantage of rapid decoding, thus achieve accurate consumption intention inference. The experimental results on real datasets show that compared with the baseline models, the AUC of tourist behaviour and consumption intention inference achieve at least 2.8% and 3.5% improvement respectively, which demonstrate the excellent inference performance.

Acknowledgements

This work is supported by the Jiaying University 'The Perspective of Rural Revitalization on Social Innovation Model in Tourism Enterprises' (No. 2022WRC01).

Declarations

All authors declare that they have no conflicts of interest.

References

An, S., Kim, W., Lee, B. and Suh, J. (2022) 'A study on the tourism-related information consumption process of tourists on social networking sites', *Sustainability*, Vol. 14, No. 7, pp.39–50.

Bai, S. and Han, F. (2020) 'Tourist behavior recognition through scenic spot image retrieval based on image processing', *Traitement du Signal*, Vol. 37, No. 4, pp.71–84.

Choe, J.Y.J. and Kim, S.S. (2018) 'Effects of tourists' local food consumption value on attitude, food destination image, and behavioural intention', *International Journal of Hospitality Management*, Vol. 71, pp.1–10.

Fang, F., Li, L., Zhu, H. and Lim, J.-H. (2019) 'Combining faster R-CNN and model-driven clustering for elongated object detection', *IEEE Transactions on Image Processing*, Vol. 29, pp.2052–2065.

Gregoriades, A., Pampaka, M., Herodotou, H. and Christodoulou, E. (2023) 'Explaining tourist revisit intention using natural language processing and classification techniques', *Journal of Big Data*, Vol. 10, No. 1, pp.60–82.

Kang, J., Guo, X., Fang, L., Wang, X. and Fan, Z. (2022) 'Integration of internet search data to predict tourism trends using spatial-temporal XGBoost composite model', *International Journal of Geographical Information Science*, Vol. 36, No. 2, pp.236–252.

Koo, C., Joun, Y., Han, H. and Chung, N. (2016) 'A structural model for destination travel intention as a media exposure: belief-desire-intention model perspective', *International Journal of Contemporary Hospitality Management*, Vol. 28, No. 7, pp.1338–1360.

Li, J. and Cao, B. (2022) 'Study on tourism consumer behaviour and countermeasures based on big data', *Computational Intelligence and Neuroscience*, Vol. 4, No. 1, pp.61–78.

Li, Y., Xu, S., Zhang, B., Cao, X., Gao, P. and Guo, G. (2022) 'Q-vit: accurate and fully quantized low-bit vision transformer', *Advances in Neural Information Processing Systems*, Vol. 35, pp.34451–34463.

Liu, N. and Hu, D. (2025) 'The design of consumer behaviour prediction and optimisation model by integrating DQN and LSTM', *PLOS One*, Vol. 20, No. 7, pp.32–48.

Liu, W., Zhou, B., Wang, Z., Yu, G. and Yang, S. (2023) 'FPPNet: a fixed-perspective-perception module for small object detection based on background difference', *IEEE Sensors Journal*, Vol. 23, No. 10, pp.11057–11069.

Mou, T. and Wang, H. (2025) 'Online comments of tourist attractions combining artificial intelligence text mining model and attention mechanism', *Scientific Reports*, Vol. 15, No. 1, pp.10–23.

Rezapouraghdam, H., Akhshik, A. and Ramkissoon, H. (2023) 'Application of machine learning to predict visitors' green behaviour in marine protected areas: evidence from Cyprus', *Journal of Sustainable Tourism*, Vol. 31, No. 11, pp.2479–2505.

Rong, J., Hao, H. and Xu, W. (2024) 'Big data intelligent tourism management platform design based on abnormal behaviour identification', *Intelligent Systems with Applications*, Vol. 21, pp.20–35.

Senbeto, D.L. and Hon, A.H. (2020) 'The impacts of social and economic crises on tourist behaviour and expenditure: an evolutionary approach', *Current Issues in Tourism*, Vol. 23, No. 6, pp.740–755.

Shao, X. and Kim, C.S. (2020) 'Multi-step short-term power consumption forecasting using multi-channel LSTM with time location considering customer behaviour', *IEEE Access*, Vol. 8, pp.125263–125273.

Sharma, M.P., Meena, U. and Sharma, G.K. (2022) 'Intelligent data analysis using optimised 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, W. (2025) 'Application of deep learning in consumer purchase intention prediction', *Service Oriented Computing and Applications*, Vol. 2, pp.1–15.

Trebing, K., Stańczyk, T. and Mehrkanoon, S. (2021) 'SmaAt-UNet: precipitation nowcasting using a small attention-UNet architecture', *Pattern Recognition Letters*, Vol. 145, pp.178–186.

Wu, D.C., Zhong, S., Wu, J. and Song, H. (2025) 'Tourism and hospitality forecasting with big data: a systematic review of the literature', *Journal of Hospitality & Tourism Research*, Vol. 49, No. 3, pp.615–634.

Yao, T., Li, Y., Pan, Y., Wang, Y., Zhang, X.-P. and Mei, T. (2023) 'Dual vision transformer', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 45, No. 9, pp.10870–10882.

Ye, Z. and Huang, X. (2022) 'Adoption of a deep learning-based neural network model in the psychological behaviour analysis of resident tourism consumption', *Frontiers in Public Health*, Vol. 10, pp.99–112.

Yu, W., Liao, X., Ji, S. and Cui, F. (2025) 'Green rewards vs. non-green rewards? The impact of hotel marketing incentives on guests' green consumption intentions', *Journal of Sustainable Tourism*, Vol. 33, No. 8, pp.1534–1552.

Zhang, Q., Zhang, M., Chen, T., Sun, Z., Ma, Y. and Yu, B. (2019) 'Recent advances in convolutional neural network acceleration', *Neurocomputing*, Vol. 323, pp.37–51.

Zhang, X., Cheng, M. and Wu, D.C. (2025) 'Daily tourism demand forecasting and tourists' search behaviour analysis: a deep learning approach', *International Journal of Machine Learning and Cybernetics*, Vol. 16, No. 10, pp.7133–7146.

Zhang, X., Song, Y., Song, T., Yang, D., Ye, Y., Zhou, J. and Zhang, L. (2024) 'LDConv: linear deformable convolution for improving convolutional neural networks', *Image and Vision Computing*, Vol. 149, pp.105–119.

Zhao, L. and Zhang, Z. (2024) 'An improved pooling method for convolutional neural networks', *Scientific Reports*, Vol. 14, No. 1, pp.15–24.