Journal of Network Intelligence ©2026 ISSN 2414-8105
Taiwan Ubiquitous Information Volume **11**, Number **1**, February **2026**

Multimodal City Image Short Video Recommendation Based on Graph Neural Network

Jia-Xin Zhao[1,*]

[1]College of Humanities and Management,
Xi'an Traffic Engineering Institute,
Xi'an Shaanxi, 710300, P. R. China
ZJX436095933@163.com

Xiao-Tong Wang[2]

[2]Yana Royal Polytechnic University,
Chiang Mai 50000, Thailand
xt8950@163.com

[*]Corresponding author: Jia-Xin Zhao

Received October 5, 2024, revised January 11, 2025, accepted April 25, 2025.

Abstract. *Traditional short video recommendation systems for city image are difficult to effectively capture the dynamic changes of user interests and integrate multimodal information, resulting in insufficient accuracy and diversity of recommendations. With the rapid development of short video platforms and the increasing demand for city promotion, there is an increasingly urgent need for systems that can handle large-scale data and achieve personalised recommendations. To solve these problems, a spatio-temporal aware multi-interest fusion recommendation model (TSMR) based on graph neural networks is proposed. Firstly, with the spatio-temporal-aware self-supervised graph learning module (TSGL), the model is able to capture the dynamic evolution of user interests in real time, which improves the timeliness of recommendations. Second, the Adaptive Multimodal Multi-Interest Extraction Module (AMMIE) effectively fuses the visual, audio and textual features of videos to enhance the representation of diverse user interests. Finally, the Hierarchical Sampling and Aggregation Algorithm (HSAA) significantly improves the computational efficiency of the model on large-scale data. The experimental results show that, compared with the best baseline method, TSMR model has improved the MRR and NDCG@10 by 7.16% and 5.93% respectively, and NDCG@10 has improved by 9.94% and 9.47% in dealing with the cold start of new users and new videos. These improvements provide strong technical support for the accurate recommendation and effective dissemination of short video of city image.*

Keywords: short city image video; graph neural network; multimodal recommendation; spatio-temporal awareness; multi-interest modelling.

1. **Introduction.** With the rapid development and popularity of short video platforms, city image publicity and promotion has entered a new era. With its intuitive, vivid and easy-to-communicate characteristics, short videos have become an important medium to display city characteristics and attract tourists and investment. However, in the face of a huge amount of short video content on city image, how to recommend the appropriate videos to potential audiences accurately has become a key problem that needs to be solved [1-3].

Traditional video recommendation methods, such as collaborative filtering and content-based recommendation, face a number of challenges when dealing with highly personalised and time-sensitive content such as short city image videos [4]. Firstly, users' interests in cities are often diverse and dynamically changing, and may be affected by factors such as seasons and hot events. Second, short city image videos usually contain rich multimodal information (visual, audio, and text) [5, 6], and the effective integration of this information is crucial for accurately grasping the video content. Finally, city image promotion often requires balance and diversity, recommending popular attractions as well as discovering niche but distinctive content.

In recent years, deep learning techniques, especially graph neural networks (GNNs), have made significant progress in the field of recommender systems [7-9]. GNNs are able to effectively capture the complex patterns of user-video interactions, providing new possibilities for personalised recommendations. However, existing GNN-based methods still have limitations in handling dynamically changing user interests and fusing multimodal information. In addition, how to train and reason efficiently on large-scale datasets is also an important challenge.

1.1. Related work.

The evolution of video recommender systems has witnessed innovations from traditional methods to deep learning techniques, which continue to drive the advancement of recommendation algorithms.

Traditional video recommendation methods mainly rely on collaborative filtering and content-based recommendations. Lin and Shyu [10] proposed an association rule-based video recommendation algorithm that generates recommendation lists by mining implicit patterns in users' viewing history. It was found that the method performs well when dealing with sparse data, but it is difficult to capture the dynamic changes in user interests, especially in the rapidly updating short video scenarios. Bokde et al. [11] designed a hybrid model combining matrix decomposition and factoriser to solve the cold-start problem. However, this approach is computationally inefficient when dealing with large-scale sparse data and is difficult to effectively exploit the multimodal features of videos.

Current mainstream approaches for multimodal video recommendation tend to utilise deep learning techniques to improve performance. Fang et al. [12] developed a sequential recommendation model based on an attention mechanism that can effectively capture both long-term and short-term changes in user interests. Although the model has made significant progress in temporal sequence modelling, it still has limitations when dealing with multimodal data. Tao et al. [13] proposed a multimodal graph convolutional network framework to learn richer representations by constructing user-video interaction graphs. While this approach makes progress in exploiting multimodal information, it still does not fully explore the complex relationships between modalities. Sang et al. [14] devised a multimodal recommendation model incorporating self-attention and graph attention, which demonstrated superior performance in modelling user interests and video features. However, the computational efficiency and scalability of the model in dealing with large-scale sparse data still need to be improved.

Graph neural network techniques show great potential in the field of multimodal video recommendation. Tao et al. [15] proposed a self-supervised graph learning framework to enhance the generalisation of node representations through contrast learning. It is shown that although this approach has made significant progress in improving recommendation performance, it still faces challenges in dealing with dynamically changing user interests, especially in short-video scenarios with diverse content and rapidly changing user preferences. Ma et al. [16] developed a transformer-based multi-interest extraction model that is capable of capturing diversified interest preferences from users' historical behaviours.

However, how to effectively fuse multimodal information as well as balance the accuracy and diversity of recommendations still requires in-depth research, especially when dealing with large-scale and highly personalised short video data.

1.2. Motivation and contribution.

Existing multimodal video recommendation approaches still face significant challenges in handling dynamic user interests, fusing multimodal information and coping with large-scale personalised data. These problems are especially prominent in the rapidly changing and content-rich domain of short city image videos. To address the above problems, a spatio-temporally aware multimodal short city image video recommendation model fusing self-supervised graph learning and multi-interest extraction is proposed. The main innovations and contributions of this work include:

(1) A spatio-temporally aware self-supervised graph learning mechanism is proposed to address the limitations of Tao et al. [15] in dealing with dynamically changing user interests. The mechanism achieves real-time capture and modelling of user interest evolution by introducing temporal coding and dynamic graph update strategies. This innovation effectively addresses the shortcomings of existing models in dealing with rapidly changing user preferences, and especially excels in scenarios such as short city image videos with diverse contents and rapidly changing user interests.

(2) To address the challenges of Ma et al. [16] in multimodal information fusion and diverse interest modelling, an adaptive multimodal multi-interest extraction module is designed. The module better captures the complex interactions between visual, audio and textual information in short city image videos by dynamically adjusting the importance of different modal features. Meanwhile, an improved attention mechanism is used to extract diverse interest representations from users' historical behaviours, which effectively improves the accuracy and diversity of recommendations.

(3) In order to solve the efficiency problem of large-scale personalised data processing, a hierarchical sampling and aggregation algorithm based on graph neural network is developed. The algorithm significantly improves the computational efficiency and scalability of the model in processing massive city image short video data through adaptive neighbourhood sampling and multi-layer information aggregation, while maintaining the level of personalisation of recommendations.

Through the above innovations, we not only targeted to solve the deficiencies of the existing methods in dynamic interest modelling, multimodal information fusion and large-scale data processing, but also provide new research ideas and technical solutions in the field of city image short video recommendation. These improvements will significantly enhance the performance and user experience of the recommendation system and provide strong technical support for the effective communication and promotion of city image.

2. Model architecture.

2.1. Problem definition.

In this section, a formal definition of the city image short video recommendation task is provided to lay the foundation for the subsequent description of the model architecture.

Let $\mathcal{U} = \{u_1, u_2, \ldots, u_M\}$ denote the set of users, where M is the total number of users; $\mathcal{V} = \{v_1, v_2, \ldots, v_N\}$ denotes the set of short city image videos, where N is the total number of videos. Each video v_i contains multimodal features defined as $v_i = \{v_i^{\text{visual}}, v_i^{\text{audio}}, v_i^{\text{text}}\}$, which denote the visual, audio and textual features, respectively.

The history of a user's interactions with a video can be represented as a two-part graph $\mathcal{G} = (\mathcal{U}, \mathcal{V}, \mathcal{E})$ where $\mathcal{E} \subseteq \mathcal{U} \times \mathcal{V}$ denotes the set of edges. Each interaction $e_{ij} =$

$(u_i, v_j, t_{ij}) \in \mathcal{E}$ contains the timestamp t_{ij}, which is used to capture dynamic changes in user interest.

To model the diverse interests of users, we define the set of interests of user u_i as $\mathcal{I}_i = \{I_{i1}, I_{i2}, \ldots, I_{iK}\}$, where K is a predefined number of interests. Each interest I_{ik} is a vector representing the user's preference in a particular aspect.

Considering the spatio-temporal characteristics of short city image videos, we introduce a temporal encoding function $\phi(t)$ to map the timestamp t to a continuous time representation space [17, 18]. Furthermore, we define the dynamic graph update strategy $\mathcal{F}(\mathcal{G}, t)$ for updating the graph structure $\mathcal{G}$ at time t.

Based on the above definition, the city image short video recommendation task can be formalised as follows: given a user u_i, time t, a historical interaction graph $\mathcal{G}$ and a set of candidate videos $\mathcal{V}_c \subset \mathcal{V}$, predict the user's interest score s_{ij} at time t in respect of each video $v_j \in \mathcal{V}_c$. The mathematical expression is as follows:

$$s_{ij} = f(u_i, v_j, t, \mathcal{G}, \mathcal{I}_i) \tag{1}$$

where $f(\cdot)$ is the recommendation model we are going to learn. This model needs to take into account the diverse interests of users, the multimodal characteristics of videos, and the spatio-temporal dynamics in order to achieve accurate personalised recommendation.

The core objective of this study is to design an efficient recommendation model $f(\cdot)$ by incorporating spatio-temporal aware self-supervised graph learning, adaptive multimodal multi-interest extraction and hierarchical sampling aggregation algorithms in order to improve the accuracy, diversity and timeliness of short city image video recommendations. In the next sections, we will detail the general framework of the model and the design of each key module.

2.2. **General framework of the model.** The overall framework of the proposed Time-aware Self-supervised Multi-interest Recommendation (TSMR) model for the city image short video recommendation task is shown in Figure 1. The TSMR model aims to address the challenges of dynamic user interest modeling, multimodal information fusion and large-scale data processing in city image short video recommendation.

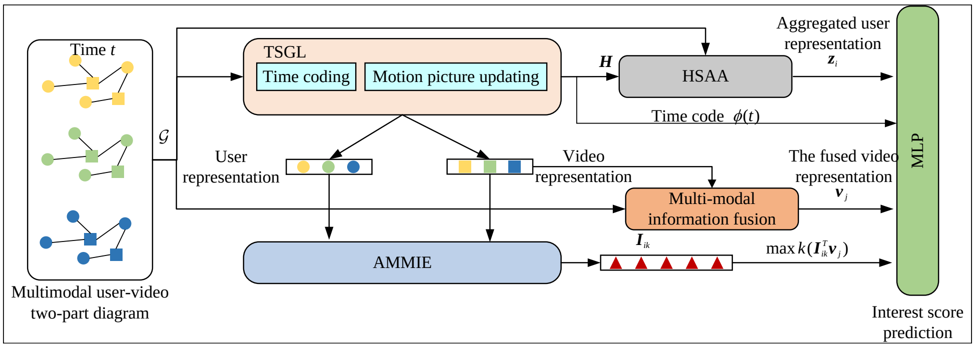

Figure 1. The TSMR model

The TSMR model consists of three core modules: the Temporal-aware Self-supervised Graph Learning (TSGL) module, the Adaptive Multimodal Multi-interest Extraction (AMMIE) module, and the Hierarchical Sampling Aggregation Algorithm (HSAA). These three modules work in tandem to achieve dynamic modelling of user interests, effective fusion of multimodal information and efficient processing of large-scale data.

The TSGL module incorporates temporal information into the graph structure by introducing temporal encoding and dynamic graph updating strategies [19, 20]. Given a user-video interaction graph $\mathcal{G}$ and time t, the TSGL module first applies a temporal encoding function $\phi(t)$ to generate a temporal representation, and then updates the graph structure using a dynamic graph update strategy $\mathcal{F}(\mathcal{G}, t)$. The updated graph structure is fed into a self-supervised learning task to generate spatio-temporally aware node representations:

$$\mathbf{H} = \text{TSGL}(\mathcal{G}, t) \tag{2}$$

where $\mathbf{H} \in \mathbb{R}^{(M+N)\times d}$ is the learned node representation matrix and d is the dimension of the representation vector.

The AMMIE module is responsible for fusing multimodal information and extracting the user's diverse interests. For the video v_j, the AMMIE module first fuses its multimodal features:

$$\mathbf{v}_j = \text{AMMIE}_{\text{fusion}}(v_j^{\text{visual}}, v_j^{\text{audio}}, v_j^{\text{text}}) \tag{3}$$

The AMMIE module then extracts K interest representations from the user's historical interactions using a dynamic attention mechanism:

$$\mathcal{I}_i = \text{AMMIE}_{\text{interest}}(\mathbf{H}_i, \{\mathbf{v}_j \mid (u_i, v_j, t_{ij}) \in \mathcal{E}\}) \tag{4}$$

where $\mathbf{H}_i$ is the node representation of user u_i.

The inputs to the HSAA include the node representation matrix $\mathbf{H}$, the user-video interaction graph $\mathcal{G}$ and the number of layers L. The $\mathbf{H}$ contains all the initial representations of users and videos and is generated by the TSGL. $\mathcal{G}$ is the graph structure we mentioned before. It contains information about the interaction between the user and the video. L is a hyperparameter that determines the depth of the graph convolution.

The HSAA module enables efficient processing of large-scale graph data through adaptive neighbour sampling and multi-layer information aggregation. Given the centre node u_i and the number of layers L, the HSAA module performs the following operations:

$$\mathbf{z}_i = \text{HSAA}(\mathbf{H}_i, \mathcal{G}, L) \tag{5}$$

where $\mathbf{z}_i$ is the aggregated user representation.

Ultimately, the TSMR model calculates the interest score of a user u_i in a video v_j at time t by taking into account the diverse interests of the user, the multimodal characteristics of the video, and the spatio-temporal dynamics:

$$s_{ij} = f(u_i, v_j, t, \mathcal{G}, \mathcal{I}_i) = \text{MLP}([\mathbf{z}_i; \mathbf{v}_j; \phi(t); \max_k(\mathbf{I}_{ik}^T \mathbf{v}_j)]) \tag{6}$$

where MLP$(\cdot)$ denotes a multilayer perceptual machine, $[\cdot;\cdot]$ denotes a vector splicing operation, and $\max_k(\mathbf{I}_{ik}^T \mathbf{v}_j)$ selects the most relevant interest to the video.

With this design, the TSMR model is able to effectively capture the dynamic changes of user interests, make full use of multimodal information, and achieve efficient recommendation on large-scale data. Next, we will introduce the specific implementation of each core module in detail.

2.3. **TSGL module.** TSGL is a core component of the TSMR model designed to capture the dynamic evolution of user interests and the temporal relationships of video content, as shown in Figure 2. The TSGL module generates spatio-temporal-aware node representations by fusing temporal information and graph structure, laying the foundation for subsequent multi-interest extraction and recommendation tasks.

The input to the TSGL module consists of a user-video interaction graph $\mathcal{G}$, temporal information t and multimodal features $\{v^{\text{visual}}, v^{\text{audio}}, v^{\text{text}}\}$. The module first applies the temporal encoding function $\phi(t)$ to map discrete timestamps to a continuous vector space, then updates the graph structure using the dynamic graph update strategy, and finally optimises the node representations by self-supervised learning objectives.

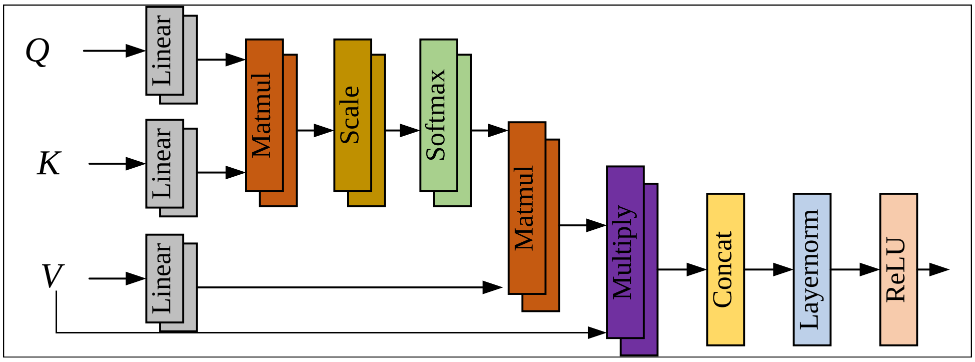

Figure 2. TSGL module

The time encoding function $\phi(t)$ uses a sinusoidal positional encoding method to map the timestamp t into a d dimensional vector:

$$\phi(t)_i = \begin{cases} \sin(t/10000^{2i/d}), & \text{if } i \text{ is even} \\ \cos(t/10000^{2i/d}), & \text{if } i \text{ is odd} \end{cases} \tag{7}$$

where i denotes the i-th dimension of the vector. This encoding method maintains the relative order of time and captures periodic patterns at different time scales.

The dynamic graph update strategy $\mathcal{F}(\mathcal{G}, t)$ adjusts the edge weights through a time decay function:

$$w_{ij}^t = w_{ij} \cdot \exp(-\lambda(t - t_{ij})) \tag{8}$$

where w_{ij} is the original weight of edge (i, j), t_{ij} is the time when the interaction occurs, and λ is the decay coefficient. This strategy makes recent interactions have higher weights, thus reflecting the dynamic changes in user interests.

The TSGL module uses contrast learning as a self-supervised objective to learn spatio-temporally aware node representations by maximising the mutual information of the same node represented in different views. Specifically, for node v, we construct two views $\tilde{v}_1$ and $\tilde{v}_2$, and obtain their representations $\mathbf{h}_1$ and $\mathbf{h}_2$ via GCN. The loss function for contrast learning is defined as:

$$\mathcal{L}_{\text{contrast}} = -\log \frac{\exp(\mathbf{h}_1^T \mathbf{h}_2/\tau)}{\sum_{v' \in \mathcal{N}(v)} \exp(\mathbf{h}_1^T \mathbf{h}_{v'}/\tau)} \tag{9}$$

where τ is the temperature parameter and $\mathcal{N}(v)$ is the set of negative samples of node v.

By optimising this self-supervised learning objective, the TSGL module is able to generate spatio-temporally aware node representation matrices $\mathbf{H} \in \mathbb{R}^{(M+N)\times d}$, which contain the representations of both the user and video nodes. The innovation of the TSGL module is that it effectively incorporates temporal information into the graph learning process, capturing the evolving patterns of user interest through temporal coding and dynamic graph updating strategies.

2.4. **AMMIE module.** AMMIE is designed to capture the diverse interests of users and to fuse multimodal information. The AMMIE module employs a multi-head structure based on the attention mechanism to efficiently extract and represent multiple interests of users. The inputs to the AMMIE module include the query matrix $\mathbf{Q}$, the key matrix $\mathbf{K}$, and the value matrix $\mathbf{V}$, which are obtained by linearly transforming the original features.

2.4.1. *Multimodal feature fusion.* Multimodal feature fusion occurs before generating the $\mathbf{Q}, \mathbf{K}$, and $\mathbf{V}$ matrices in the following manner:

$$\mathbf{v}_j = \sum_{m \in \{\text{visual, audio, text}\}} \alpha_m \cdot \mathbf{W}_m v_j^m \tag{10}$$

where α_m is the adaptive weights computed by the attention mechanism and $\mathbf{W}_m$ is the transformation matrix of modality m.

2.4.2. *Multi-attention mechanisms.* The multi-head attention mechanism allows the model to simultaneously attend to different representation subspaces of the input. For each attention head k, we perform the following steps:

(1) Calculate the attention score:

$$S_k = \text{matmul}(Q_k, K_k^T) \tag{11}$$

(2) Scaling and softmax:

$$A_k = \text{softmax}\left(\frac{S_k}{\sqrt{d_k}}\right) \tag{12}$$

where d_k is the dimension of the key vector.

(3) Weighted aggregation:

$$\mathbf{I}_{ik} = \text{matmul}(A_k, V_k) \tag{13}$$

where $\mathbf{I}_{ik}$ denotes the k-th interest representation of user u_i.

2.4.3. *Multi-interest representation learning.* Multi-interest representation learning sets the individual interest representations $\mathbf{I}_{ik}$ into a final interest representation $\mathbf{I}_{\text{final}}$ by the following steps:

(1) Interest aggregation:

$$\mathbf{I}_{\text{concat}} = \text{concat}(\mathbf{I}_{i1}, \mathbf{I}_{i2}, \ldots, \mathbf{I}_{iK}) \tag{14}$$

where K is the number of attention heads.

(2) Linear transformation:

$$\mathbf{I}_{\text{linear}} = \mathbf{W}_O \mathbf{I}_{\text{concat}} \tag{15}$$

where $\mathbf{W}_O$ is the learnable parameter matrix.

(3) Normalisation and activation:

$$\mathbf{I}_{\text{final}} = \text{ReLU}(\text{LayerNorm}(\mathbf{I}_{\text{linear}})) \tag{16}$$

The $\mathbf{I}_{ik}$ is the intermediate representation, representing the k-th specific interest of the user u_i. Each $\mathbf{I}_{ik}$ captures a specific aspect of the user's interest. The $\mathbf{I}_{\text{final}}$ is the final multi-interest representation, which combines information from all $\mathbf{I}_{ik}$. Through splicing, linear transformation, normalisation and nonlinear activation, $\mathbf{I}_{\text{final}}$ not only preserves the information of each specific interest, but also captures the interrelationships among them.

Through the mechanism of multi-attention, the AMMIE module is able to capture multiple aspects of a user's interests ($\mathbf{I}_{ik}$) simultaneously. Through multi-interest representation learning, the AMMIE module effectively integrates these specific interests ($\mathbf{I}_{ik}$) into a comprehensive interest representation ($\mathbf{I}_{\text{final}}$). By adaptively fusing multimodal features, the AMMIE module is able to take full advantage of the rich information of video content.

This design not only improves the model's ability to capture complex user interests, but also enhances the interpretability of the recommender system. In the city image short video recommendation task, the AMMIE module is able to efficiently extract diverse user interests and synthesise these interests into a robust representation, thus providing more accurate and personalised recommendation results.

2.5. **HSAA algorithm.** HSAA aims to efficiently process large-scale graph-structured data and generate high-quality user representations. The HSAA module overcomes the computational efficiency and representation capability problems of traditional graph neural networks in processing large-scale graph data through an innovative sampling strategy and a multi-layer information aggregation mechanism. The inputs to the HSAA module include a node representation matrix that is generated by the TSGL module $\mathbf{H} \in \mathbb{R}^{(M+N)\times d}$, a user-video interaction graph $\mathcal{G}$, and a predefined number of layers L.

2.5.1. *Adaptive neighbour sampling.* Adaptive neighbour sampling is the first innovation of HSAA, which aims to dynamically adjust the number of samples according to the importance of a node. For user node u_i, the sampling function at the l-th layer is defined as:

$$\mathcal{N}_l(u_i) = \mathrm{Sample}(\mathcal{N}(u_i), \min(|\mathcal{N}(u_i)|, \alpha \cdot d_i^{\beta})) \tag{17}$$

where $\mathcal{N}(u_i)$ is the set of neighbours of u_i, d_i is the degree of u_i, and α and β are adjustable parameters. This sampling strategy ensures that the height nodes are able to sample more neighbours and thus retain more structural information.

To further improve the efficiency and quality of sampling, we introduce an importance-based sampling mechanism. For each neighbour node v_j, we calculate its importance score:

$$s_{ij} = f(\mathbf{h}_i, \mathbf{h}_j) \tag{18}$$

where f is a learnable scoring function, $\mathbf{h}_i$ and $\mathbf{h}_j$ are representations of nodes u_i and v_j, respectively. The sampling probability is then normalised by the importance score:

$$p_{ij} = \frac{\exp(s_{ij})}{\sum_{k\in\mathcal{N}(u_i)} \exp(s_{ik})} \tag{19}$$

This adaptive sampling strategy not only improves the computational efficiency, but also retains the most valuable information in the graph, which is particularly effective for dealing with large-scale short video recommendation scenarios of city images.

2.5.2. *Multi-layer information aggregation.* Multi-layer information aggregation is the second innovation of HSAA, which aims to capture complex structural features through multi-layer information transfer. At each layer, HSAA updates the node representation by aggregating the neighbour information. The aggregation process at the l-th layer can be represented as follows:

$$\mathbf{h}_i^l = \sigma(\mathbf{W}^l \cdot \mathrm{AGGREGATE}(\{\mathbf{h}_j^{l-1}, \forall j \in \mathcal{N}_l(u_i)\})) \tag{20}$$

where $\mathbf{h}_i^l$ is the representation of node u_i at the l-th layer, $\mathbf{W}^l$ is the weight matrix at the l-th layer, σ is a nonlinear activation function, and AGGREGATE is an optional aggregation function.

Considering that the context of this study is city image short video recommendation, we need an aggregation function that can both capture complex interaction patterns and handle multimodal information. Based on these needs, we choose Attention Aggregator as the AGGREGATE function of the HSAA module. Define the AGGREGATE function as follows:

$$\mathrm{AGGREGATE}(\{\mathbf{h}_j^{l-1}, \forall j \in \mathcal{N}_l(u_i)\}) = \sum_{j\in\mathcal{N}_l(u_i)} \alpha_{ij}\mathbf{h}_j^{l-1} \tag{21}$$

where the attention weight α_{ij} is calculated by the following:

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k\in\mathcal{N}_l(u_i)} \exp(e_{ik})} \tag{22}$$

$$e_{ij} = \mathbf{q}^T \tanh(\mathbf{W}_1\mathbf{h}_i^{l-1} + \mathbf{W}_2\mathbf{h}_j^{l-1}) \tag{23}$$

where $\mathbf{W}_1, \mathbf{W}_2$ are learnable weight matrices and $\mathbf{q}$ is a learnable query vector.

The reasons for choosing the attention aggregation function are as follows:

(1) **Dynamic Importance**: the attention mechanism can dynamically assign importance based on the representation of the current node and neighbouring nodes, which is particularly important for capturing changes in user interest and relevance of video content.

(2) **Handling heterogeneity**: In city image short video recommendation, the user-video interaction graph is highly heterogeneous. The attention mechanism can effectively handle this heterogeneity by assigning different weights to different types of neighbour nodes.

(3) **Multimodal information fusion**: the attention mechanism can adaptively fuse information from different modalities (e.g., visual, audio, text), which is consistent with our goal of multimodal feature fusion.

(4) **Interpretability**: by analysing the attention weights, we can get some explanation of the model's decisions and understand which neighbours contribute more to the representation of the current node.

(5) **Nonlinear representation**: the tanh activation function introduces nonlinearity, allowing the model to learn more complex interaction patterns.

Integrating the attention aggregation function into the HSAA module, the nodes in the l-th layer indicate that the update formula becomes:

$$\mathbf{h}_i^l = \sigma\left(\mathbf{W}^l \cdot \sum_{j\in\mathcal{N}_l(u_i)} \alpha_{ij}\mathbf{h}_j^{l-1}\right) \tag{24}$$

This design allows the HSAA module to more accurately capture the complex relationship between user interest and video content, while taking into account the multimodal nature of short urban image videos. It allows the model to dynamically adjust the importance of each neighbour according to the current context when aggregating neighbour information, thus generating more expressive and targeted node representations.

3. Experimentation and evaluation.

3.1. **Dataset.** In order to comprehensively evaluate the performance of the TSMR model in the city image short video recommendation task, we constructed a large-scale multimodal dataset called CityVideoRec. The dataset contains city-related content from several major short video platforms, covering image display videos of 50 Chinese cities. The statistical information of the CityVideoRec dataset is shown in Table 1.

Table 1. Statistical information on the CityVideoRec dataset

Features	**Numerical value**
Number of users	1,000,000
Number of videos	500,000
Record of interactions	10,000,000
Average video duration	45 s
Time span	2022.1.1–2022.12.31

Each video in the CityVideoRec dataset contains the following multimodal features:

(1) **Visual features**: 2048-dimensional feature vectors extracted using the pre-trained ResNet-152 model.
(2) **Audio features**: 128-dimensional feature vectors extracted using the VGGish model.
(3) **Text features**: video titles and descriptions are encoded using the BERT model to obtain 768-dimensional feature vectors.

The user-video interaction graph $\mathcal{G}$ is constructed based on the user's viewing history, where the weights of the edges e_{ij} are calculated based on the viewing duration and interaction behaviours (e.g., liking, commenting, and sharing) of the user u_i on the video v_j. The time information t records the specific timestamp of each interaction. In order to validate the ability of the TSMR model in handling the cold-start problem and capturing the dynamic evolution of user interests, we purposely include a certain percentage of new users and new videos in the dataset. Specifically, we randomly select 10% of users and videos as cold-start samples, which only appear in the test set and have no training history.

3.2. **Evaluation metrics.** In order to comprehensively evaluate the performance of the TSMR model in the task of short city image video recommendation, a series of evaluation metrics covering different aspects are used. These metrics include accuracy, diversity, novelty and temporal relevance.

Firstly, we use the Mean Reverse Rank (MRR) as the main accuracy metric. The MRR is calculated as follows:

$$\mathrm{MRR} = \frac{1}{|U|}\sum_{u\in U}\frac{1}{rank_u} \tag{25}$$

where $|U|$ is the size of the set of test users and $rank_u$ is the rank of the first relevant video in the recommendation list.

Another important accuracy metric is the Normalised Discount Cumulative Gain (NDCG@K), which takes into account the position and relevance of the items in the recommended list:

$$\mathrm{NDCG@}K = \frac{1}{|U|}\sum_{u\in U}\frac{\mathrm{DCG@}K_u}{\mathrm{IDCG@}K_u} \tag{26}$$

where $\mathrm{DCG@}K_u = \sum_{i=1}^{K}\frac{2^{rel_i}-1}{\log_2(i+1)}$, rel_i is the relevance score of the i-th recommended item.

In order to evaluate the diversity of recommendation results, we introduce the Internal List Diversity (ILD@K) metric:

$$\mathrm{ILD@}K = \frac{1}{|U|}\sum_{u\in U}\frac{\sum_{i=1}^{K}\sum_{j=i+1}^{K}d(v_i,v_j)}{K(K-1)/2} \tag{27}$$

where $d(v_i, v_j)$ is the distance between videos v_i and v_j, computed based on their multimodal features.

Novelty is measured using Average Popularity (AP@K):

$$\mathrm{AP@}K = \frac{1}{|U|}\sum_{u\in U}\frac{1}{K}\sum_{i=1}^{K}\log_2 \mathrm{pop}(v_i) \tag{28}$$

where $\mathrm{pop}(v_i)$ is the popularity of video v_i.

Finally, we design the Temporal Relevance Score (TRS@K) to assess the temporal dynamics:

$$\mathrm{TRS@}K = \frac{1}{|U|}\sum_{u\in U}\frac{1}{K}\sum_{i=1}^{K}e^{-\lambda(t_{\mathrm{rec}}-t_i)} \tag{29}$$

where t_{rec} is the recommendation time, t_i is the upload time of video v_i, and λ is the time decay factor.

3.3. **Baseline methods.** We selected several representative baseline methods for comparison:

- **BPR** [21]: A Bayesian personalised ranking model using matrix decomposition.
- **NCF** [22]: Neural collaborative filtering combining MLP and matrix decomposition.
- **NGCF** [23]: Neural graph collaborative filtering capturing higher-order connectivity.
- **LightGCN** [24]: A lightweight version of NGCF.
- **MMGCN** [25]: Specifically designed for multimodal information.
- **TGN** [26]: Temporal graph networks for dynamic graph structures.
- **SLMRec** [15]: A self-supervised graph contrast learning model.
- **MrTransformer** [16]: A Transformer-based multi-interest extraction model.

3.4. **Experimental setup.** The CityVideoRec dataset is randomly divided into training, validation, and testing sets (8:1:1) using chronological division. The main hyperparameter settings are shown in Table 2.

Table 2. Main hyperparameter settings of the TSMR model

Hyperparameterisation	**Numerical value**
Embedding dimension d	64
Number of graph convolution layers L	3
Learning rate	0.001
Batch size	1024
Regularisation factor	0.0001
Attention head count K	4
Time decay factor λ	0.5
Number of sampling neighbours	20
Comparative learning temperature parameters τ	0.2

3.5. **Performance comparison.**

3.5.1. *Overall performance evaluation.* Table 3 shows the performance comparison results.

As can be observed from Table 3, the TSMR model significantly outperforms the baseline approach. Specifically, TSMR is improved by 7.16% and 5.93% on MRR and NDCG@10 compared with the best baseline. On the diversity metric ILD@10, TSMR improves by 1.44% over MrTransformer.

3.5.2. *Cold start scenario analysis.* Figure 3 illustrates the performance in cold start scenarios.

The TSMR model shows a significant advantage in handling new users and new videos. For new users, the NDCG@10 of TSMR is 9.94% higher than MrTransformer; for new videos, the improvement is 9.47%. This significant improvement is due to the TSGL module capturing temporal dynamics and the AMMIE module's multimodal fusion capability.

Table 3. Model Performance Comparison Results

Model	**MRR**	**NDCG@10**	**ILD@10**	**AP@10**	**TRS@10**
BPR	0.2134	0.2567	0.6823	5.4321	0.5678
NCF	0.2456	0.2901	0.7012	5.2109	0.5901
NGCF	0.2689	0.3145	0.7234	4.9876	0.6123
LightGCN	0.2801	0.3276	0.7345	4.8765	0.6234
MMGCN	0.2934	0.3412	0.7456	4.7654	0.6345
TGN	0.3045	0.3534	0.7567	4.6543	0.6456
SLMRec	0.3178	0.3689	0.7712	4.5321	0.6623
MrTransformer	0.3256	0.3745	0.7789	4.4987	0.6701
TSMR	**0.3489**	**0.3967**	**0.7901**	**4.3210**	**0.6901**

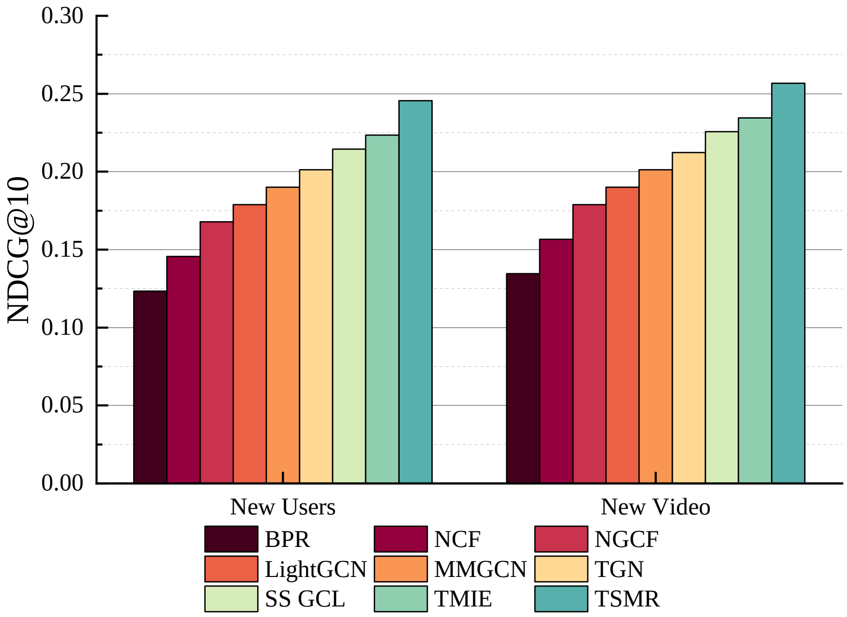

Figure 3. Cold Start Scenario NDCG@10 Performance Comparison

4. **Conclusions.** A TSMR based on graph neural network is proposed, which effectively solves the limitations of the traditional city image short video recommendation system in dealing with dynamic user interests and multimodal information fusion. By introducing spatio-temporal-aware self-supervised graph learning, the model is able to capture the evolution of user interests more accurately, which significantly improves the timeliness of recommendation. In addition, the adaptive multimodal multi-interest extraction module further enhances the ability to express diverse user interests, ensuring the diversity and personalisation of recommendation results. The following conclusions can be drawn from the experiments conducted on the CityVideoRec dataset:

(1) TSGL can significantly improve the ability of recommendation models to capture dynamic changes in user interests.

(2) The AMMIE module performs better in dealing with complex user-video interaction relationships than the traditional single interest model.

(3) HSAA effectively improves the computational efficiency of the model on large-scale data, making the TSMR model more scalable in practical applications.

(4) The TSMR model excels in dealing with the cold-start problem and provides an effective solution for recommending new users and new videos.

The experimental data mainly comes from the constructed CityVideoRec dataset, and although it covers short video content from multiple cities, the geographical nature of the data may limit the generalisation ability of the model. Future work should consider introducing more city image short video data from different countries and cultural backgrounds to verify the effectiveness of the model in cross-cultural recommendation scenarios. In addition, the incorporation of users' geographic location information and travel preferences into the model can be explored to provide more personalised and contextualised recommendation services.

Acknowledgment. This work is supported by 2023 Young and Middle aged Fund Project of Xi'an Traffic Engineering Institute (No. 2023KY-80): Research on Shaanxi Image Reporting of Mainstream Media Based on Corpus.

REFERENCES

[1] X. Cheng, J. Liu, and C. Dale, "Understanding the characteristics of internet short video sharing: A youtube-based measurement study," *IEEE Transactions on Multimedia*, vol. 15, no. 5, pp. 1184-1194, 2013.

[2] X. Chen, D. B. Valdovinos Kaye, and J. Zeng, "# PositiveEnergy Douyin: Constructing "playful patriotism" in a Chinese short-video application," *Chinese Journal of Communication*, vol. 14, no. 1, pp. 97-117, 2021.

[3] X. Cao, Z. Qu, Y. Liu, and J. Hu, "How the destination short video affects the customers' attitude: The role of narrative transportation," *Journal of Retailing and Consumer Services*, vol. 62, 102672, 2021.

[4] D. B. V. Kaye, X. Chen, and J. Zeng, "The co-evolution of two Chinese mobile short video apps: Parallel platformization of Douyin and TikTok," *Mobile Media & Communication*, vol. 9, no. 2, pp. 229-253, 2021.

[5] Y. Myagmar-Ochir, and W. Kim, "A survey of video surveillance systems in smart city," *Electronics*, vol. 12, no. 17, 3567, 2023.

[6] L. Tian, H. Wang, Y. Zhou, and C. Peng, "Video big data in smart city: Background construction and optimization for surveillance video processing," *Future Generation Computer Systems*, vol. 86, pp. 1371-1382, 2018.

[7] S. Wu, F. Sun, W. Zhang, X. Xie, and B. Cui, "Graph neural networks in recommender systems: a survey," *ACM Computing Surveys*, vol. 55, no. 5, pp. 1-37, 2022.

[8] C. Gao, Y. Zheng, N. Li, Y. Li, Y. Qin, J. Piao, Y. Quan, J. Chang, D. Jin, and X. He, "A survey of graph neural networks for recommender systems: Challenges, methods, and directions," *ACM Transactions on Recommender Systems*, vol. 1, no. 1, pp. 1-51, 2023.

[9] Y. Chu, J. Yao, C. Zhou, and H. Yang, "Graph neural networks in modern recommender systems," *Graph Neural Networks: Foundations, Frontiers, and Applications*, pp. 423-445, 2022.

[10] L. Lin, and M.-L. Shyu, "Weighted association rule mining for video semantic detection," *International Journal of Multimedia Data Engineering and Management (IJMDEM)*, vol. 1, no. 1, pp. 37-54, 2010.

[11] D. Bokde, S. Girase, and D. Mukhopadhyay, "Matrix factorization model in collaborative filtering algorithms: A survey," *Procedia Computer Science*, vol. 49, pp. 136-146, 2015.

[12] H. Fang, D. Zhang, Y. Shu, and G. Guo, "Deep learning for sequential recommendation: Algorithms, influential factors, and evaluations," *ACM Transactions on Information Systems (TOIS)*, vol. 39, no. 1, pp. 1-42, 2020.

[13] Z. Tao, Y. Wei, X. Wang, X. He, X. Huang, and T.-S. Chua, "Mgat: Multimodal graph attention network for recommendation," *Information Processing & Management*, vol. 57, no. 5, 102277, 2020.

[14] L. Sang, M. Xu, S. Qian, M. Martin, P. Li, and X. Wu, "Context-dependent propagating-based video recommendation in multimodal heterogeneous information networks," *IEEE Transactions on Multimedia*, vol. 23, pp. 2019-2032, 2020.

[15] Z. Tao, X. Liu, Y. Xia, X. Wang, L. Yang, X. Huang, and T.-S. Chua, "Self-supervised learning for multimedia recommendation," *IEEE Transactions on Multimedia*, vol. 25, pp. 5107-5116, 2022.

[16] M. Ma, P. Ren, Z. Chen, Z. Ren, H. Liang, J. Ma, and M. De Rijke, "Improving transformer-based sequential recommenders through preference editing," *ACM Transactions on Information Systems*, vol. 41, no. 3, pp. 1-24, 2023.

[17] A. Nayyar, and D. Teneketzis, "On the structure of real-time encoding and decoding functions in a multiterminal communication system," *IEEE Transactions On Information Theory*, vol. 57, no. 9, pp. 6196-6214, 2011.

[18] A. A. Lazar, and E. A. Pnevmatikakis, "Video time encoding machines," *IEEE Transactions on Neural Networks*, vol. 22, no. 3, pp. 461-473, 2011.

[19] H. Du, and N. Zhang, "Time series prediction using evolving radial basis function networks with new encoding scheme," *Neurocomputing*, vol. 71, no. 7-9, pp. 1388-1400, 2008.

[20] K. Adam, A. Scholefield, and M. Vetterli, "Sampling and reconstruction of bandlimited signals with multi-channel time encoding," *IEEE Transactions on Signal Processing*, vol. 68, pp. 1105-1119, 2020.

[21] S. Bhaggiaraj, "Ranking Prediction of Cloud Services based on BPR," *International Journal of Computer Applications*, vol. 89, no. 9, pp. 26–31, 2014.

[22] W. Chen, F. Cai, H. Chen, and M. D. Rijke, "Joint neural collaborative filtering for recommender systems," *ACM Transactions on Information Systems (TOIS)*, vol. 37, no. 4, pp. 1-30, 2019.

[23] M. Liu, J. Li, K. Liu, C. Wang, P. Peng, G. Li, Y. Cheng, G. Jia, and W. Xie, "Graph-ICF: Item-based collaborative filtering based on graph neural network," *Knowledge-Based Systems*, vol. 251, 109208, 2022.

[24] D. Mei, N. Huang, and X. Li, "Light graph convolutional collaborative filtering with multi-aspect information," *IEEE Access*, vol. 9, pp. 34433-34441, 2021.

[25] P. Yang, W. Chen, and H. Qiu, "MMGCN: Multi-modal multi-view graph convolutional networks for cancer prognosis prediction," *Computer Methods and Programs in Biomedicine*, 108400, 2024.

[26] M. Yue, H. Liu, X. Chang, L. Zhang, and T. Li, "TGN: A Temporal Graph Network for Physics Prediction," *Applied Sciences*, vol. 14, no. 2, 863, 2024.