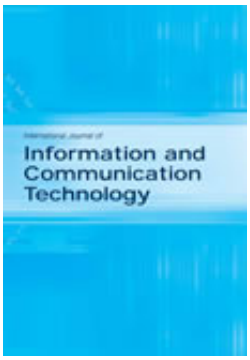
INDERSCIENCE PUBLISHERS
Linking academia, business and industry through research



International Journal of Information and Communication Technology

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

Optimisation of English learning paths in multimodal social networks using graph neural networks

Shuang Song

DOI: 10.1504/IJICT.2025.10072941

Article History:

Received:	17 June 2025
Last revised:	08 July 2025
Accepted:	08 July 2025
Published online:	08 September 2025

Optimisation of English learning paths in multimodal social networks using graph neural networks

Shuang Song

Department of Foreign Languages,
Xinxiang Institute of Engineering,
Xinxiang 453000, China
Email: ssxiaohen@126.com

Abstract: Personalised English learning based on multimodal data is in increasing demand given the fast development of information technology and social media. It is challenging to fully understand learners' behavioural traits and learning demands in multimodal social environments, nevertheless, since conventional approaches of path optimisation usually concentrate on one modality or neglect social ties. This work presents a multimodal fusion path optimisation model based on graph neural network (GNN), i.e., MMP-GENLPO, to handle the above challenges. By combining multimodal elements and applying path optimisation techniques, MMP-GENLPO efficiently increases the structural reasonableness and personalised adaptive capacity of learning paths. MMP-GENLPO has good application prospects and promotion value and verifies its efficacy and practicality in intelligent English learning recommendation by outperforming several comparative models in four key metrics by means of experimental validation on real multimodal social learning datasets.

Keywords: graph neural network; GNN; multimodal data; social networks; English learning path optimisation.

Reference to this paper should be made as follows: Song, S. (2025) 'Optimisation of English learning paths in multimodal social networks using graph neural networks', *Int. J. Information and Communication Technology*, Vol. 26, No. 32, pp.17–36.

Biographical notes: Shuang Song received her Bachelor's degree from the Henan Normal University in 2011, and Master's degree from the Henan Normal University in 2013. She is currently working as a Lecturer at the Department of Foreign Language in Xinxiang Institute of Engineering. Her research interests include English education, English translation and social networks.

1 Introduction

1.1 Context of study

Social networking sites are progressively changing from instruments for information sharing to learning support settings in the framework of growing digital education. Particularly for English students, social network multimodal content resources are altering conventional approaches of language acquisition (Lim et al., 2022). This learning

mode which consists of active participation, cross-media interaction, and context-driven learning helps to raise the immersion and involvement of education.

Though learning materials in social networks are rich, they are fragmented, heterogeneous, and semantically different. Therefore, how to extract meaningful information from them to help to construct learning paths is a crucial question in smart education today (Chen et al., 2020). While conventional recommendation systems typically rely on a single modality or fixed labels, which makes it difficult to comprehensively understand the multi-dimensional characteristics of learners, resulting in insufficient adaptability and accuracy of the recommendation path, there are differences in the behavioural habits, interest preferences, and language bases of different learners, so producing different learning effects of the same content for different users.

Graph neural network (GNN) has drawn a lot of interest in the domains of education recommendation and social behaviour modelling meanwhile thanks to its advantages in managing non-Euclidean structured data since AI technology is so widely used. GNN is appropriate for building association graphs between users and resources in social networks to accomplish personalised path planning and can fully employ node features and neighbourhood structure (Dong et al., 2023). In the English learning environment, GNN can not only better depict the dynamic learning state of learners but also investigate their possible learning path demands and increase the efficacy of recommendation techniques if it can relate to multimodal data.

Furthermore, extending the idea of learner-centred education is the suggestion of multimodal learning routes. Combining several kinds of information sources and methodically examining users' behavioural paths and cognitive preferences in social networks helps to build a more scientific, adaptable and personalised English learning path. This offers a fresh research path for the integration and development of artificial intelligence and language instruction in addition to being of pragmatic relevance for advancing the process of education informatisation.

All things considered, the building of an English learning path optimisation model based on GNN and incorporates multimodal social network information (MMP-GENLPO) has great theoretical significance and practical application prospects. It not only answers the actual needs of intelligent recommendation and tailored learning but also offers fresh technical support for investigating educational optimisation in the social network environment.

1.2 Current status of research

Social network platforms have progressively become a major arena for language learning, in which students may autonomously access materials, engage in interactive exchanges, and create different learning routes as artificial intelligence and education technology are progressively merging. Given the abundance of unstructured and multimodal data in social networks, how to extract effective and customised learning paths in this setting becomes a major challenge for intelligent education systems. Scholars have progressively built a series of technical frameworks with path generating capabilities by means of multi-dimensional exploration involving many cross-research directions including recommender systems, behavioural sequence analysis, multimodal modelling, graph representation learning, etc. so addressing this task.

Early studies mostly followed content-driven approaches for English learning materials recommendation and collaborative filtering (Chen et al., 2021). To create first

content ranking and matching recommendations, these techniques mostly rely on static similarity features between people or resources, including preference overlap rate and keyword overlap. Although this kind of method is slightly reduced in the cold-start problem, it is difficult to support continuous and more staged learning activities due of the lack of understanding the structural link between resources and the evolution of user behaviours. Consequently, collaborative-type approaches have progressively revealed their weaknesses in the English learning path optimisation problem, particularly in the context of cross-modal content and sophisticated social interaction networks, their expressive ability is very limited.

Some research has turned to sequence-based modelling approaches, using structures such as recurrent neural network (RNN) and long short-term memory (LSTM) network to capture the sequence of users' activities and their evolutionary patterns during the learning process, so improving the temporal knowledge of the model. These techniques can replicate the variations in users' access activity at several time points, therefore forecasting their future preferences and journey paths. Under some conditions, the attention mechanism helps the model to improve the emphasis on important learning nodes, hence, strengthening the rationality and goal orientation of the path. Though sequence models have great dynamic modelling capacity, their structure usually follows a linear order between resources, which makes it challenging to handle complicated branching paths and parallel tasks in the network, particularly in graphical learning environments when modelling capability is limited.

The reinforcement learning (RL)-based technique is another common research field in which the path planning problem is handled as a Markov decision process (MDP) and the model dynamically changes the recommendation strategy based on learner's state transfer and behavioural feedback. These strategies highlight the long-term optimality of learning paths and are appropriate for social learning contexts including quantifiable task goals and unambiguous feedback signals. Several researches have tried to integrate optimisation techniques including policy gradient or Q-learning to increase the total advantage of path development via environment simulation and policy iterative (Maged and Mikhail, 2020). Nevertheless, RL typically depends on accurate environment modelling and requires a lot of interaction data for training; hence, the efficiency and stability of its strategy update still face difficulties when the learning process includes multi-source information fusion and cross-modal resource coordination.

Furthermore, multimodal modelling has progressively become a research focus as the conventional text-centric recommendation approach is challenging to fit the increasingly rich modal content forms such as voice, image, video, etc. given the growing diversity of information forms in social platforms. Some work introduces modal fusion techniques, such as joint embedding and modal attention mechanism, to achieve the integration of information between different modalities at the feature layer or the decision layer, to more precisely characterise the features of learning resources and user preferences (Zhang et al., 2020). Important elements in learning route design could be, for instance, the pace of spoken language in video, the cultural background in visuals, and the semantic complexity in text. Nevertheless, among multimodalities there are notable heterogeneity and alignment challenges; so, basic fusion techniques by themselves cannot ensure the stability and robustness of path planning.

Regarding structured models, user interactions, resource connections, and behavioural paths in social networks taken together create a multi-dimensional and dynamically shifting graph structure. Graph embedding methods have been developed to extract

possible structural information across resources by means of diverse depiction of nodes and their neighbourhood connections (Cai et al., 2018). By building user-resource bipartite graphs and encoding nodes in low dimensions using graph embedding techniques such as DeepWalk, Node2Vec, and LINE, some studies have given a combined topological and semantic base representation for learning path optimisation. Nevertheless, most of these approaches depend on stationary graph structures, which are challenging to accommodate the temporal changes of learning activities, the cooperative modelling of multimodal node information, and the evolutionary traits of the structures themselves, and hence, their adaptive and generalising capacities are still limited in real social learning environments.

In essence, structural modelling ability, multimodal fusion depth, and user dynamic adaptability still have much room for development even if the present research on English learning path optimisation in social networks has attained preliminary results in recommendation strategies, behaviour prediction and content understanding. Particularly in the path generating process, how to concurrently incorporate social interactions, resource semantics, and user state changes become a difficult and crucial problem. Given its structure-aware and feature propagation powers, GNN offers fresh ideas in this context for creating a unified multimodal path optimisation model. It can not only integrate the graph structure among users, resources and behaviours, but also achieve personalised recommendation and dynamic adjustment of paths through the introduction of attention mechanism, temporal modelling and other techniques, which has progressively become the research frontier in this field.

2 Modelling multimodal social network

2.1 Multimodal social data characterisation

Multimodal social data is a collection of information modalities including text, voice, image, video, etc. and embeds the traits of users' social activities (Gandhi et al., 2023). This kind of data can more fully reflect the state characteristics of learners in the dimensions of cognition, emotion, and behaviour than single-modal data, therefore offering a rich information basis for the intelligent optimisation of English learning paths. English learners progressively access learning resources, engage in learning exchanges, and create knowledge networks on open and interactive online platforms as mobile learning and social media technology rapidly evolve. This form of learning behaviour differs greatly from the conventional classroom setting, in which users continue to produce multi-source, varied and multi-modal data while finishing courses of study.

The sources of multimodal data in real-world platform applications are quite contextual and behaviour dependent. Usually with clear topic focus and emotional expression, textual modality results from users' learning advice, comments, keyword annotations, task feedback, etc. (Zhu et al., 2024). Recorded information including speaking practice, hearing answers, voice assessment, etc. found in speech modality can represent students' success in language fluency, intonation control, and pronunciation correctness. Mostly representing the user's task execution and depth of subject knowledge, the picture modality is generated from visual records including learning punch cards, word memory maps, mind maps, etc.). Video modality emphasises more the dynamic presentation of learning activities including review of course materials,

micro-lesson narration, and scenario discussion exercises. These modalities not only differ in content but also often co-occur on the timeline, therefore forming a very complimentary multimodal social learning behavioural trajectory.

Furthermore, not disregarded should be the engagement data on social media platforms. Following, comments, likes, retweeting, and private messaging helps users create a sophisticated network of social contacts. Nestled in the learning activities, this network structure enhances the diverse and graph-structured character of the data and personalises the flow of knowledge and experience by means of their decentralisation (Siam et al., 2025). Thus, in the modelling phase, the individual characteristics of user nodes and their structural position in the social network must be considered concurrently to accomplish collaborative modelling between the data layer and the relational layer.

Direct input into the model can lead to information ambiguity and dimensional inconsistencies in the face of multimodal social data including several modalities and heterogeneous attributes. Thus, a unified fusion technique has to be developed to translate the information of several modalities into a single space for representation. Assuming the multimodal features of a node, (e.g., a user or a resource) including text characteristics $x^{(T)}$, speech features $x^{(A)}$, and picture features $x^{(I)}$, the fused feature variables can be momentarily expressed as:

$$h = x^{(T)} \| x^{(A)} \| x^{(I)} \tag{1}$$

where || signifies the variable splicing operation and the resultant variable h can be employed as a unified node representation to engage in the input propagation of GNN. Based on the preservation of the semantic characteristics of every modality, this fusion method creates a cross-modally collaborative representation space that provides the basis for node relationship modelling and feature propagation in later path optimisation.

Regarding general framework, the multimodal social data shows the following salient characteristics: first, modal richness: the data sources span several dimensions which makes the learner's state characterisation more comprehensive; second, relational mapping, the social relationships and task collaboration behaviours between users form a natural graph structure, which is suitable for the inference of the GNN-based model; third, dynamic temporal, the learning behaviour. The fourth is semantic coupling, different modalities are semantically related, e.g., the knowledge points described in the learning text may be closely related to the practice tasks in the video (Qi et al., 2021); the fifth is individual variability, the learner's background, goals, habits, and other characteristics are obviously different, and the path modelling should consider the common laws and personalised expression.

The foregoing study makes it clear that multimodal social data is not only rich and sophisticated but also rather strongly ingrained in user behaviour and social interaction. Consequently, the deep semantic information and structural elements included in these modal data should not be disregarded in the process of English learning path optimisation. The main reason structure-aware models such as GNN are presented in this work is that scientific modelling of these multimodal data can not only improve the accuracy and personalisation of learning path recommendation but also achieve an in-depth knowledge and control of the learning behaviour mechanism.

2.2 Social network structure construction method

Learners in a multimodal social English learning environment not only generate a lot of semantically differentiated learning materials but also constantly interact with others on the platform, such as liking, commenting, grouping, and retweeting, so creating a dynamically changing social network. This network must be structurally represented, that is, a graph structure that can define the multimodal learning behaviours and social links, therefore enabling effective mining of the relational structure and assistance of learning path optimisation.

Nodes in this graph can stand for several kinds of entities, including learners, learning materials, and homework (Albreiki et al., 2024). Every node carries appropriate multimodal feature variables, which modal splicing in the preceding section has consistently expressed. Conversely, edges are utilised to explain the links between nodes, that is, the social interactions among students, the reference links of resources, the dependency sequence of activities, etc. These interactions affect the propagation direction of learning paths in the graph in addition to deciding its structure.

One may officially depict the full social network graph as:

$$G = (V, E) \tag{2}$$

where E is the set of edges, signifying interactions or dependencies between entities; V is the set of nodes, therefore representing all users, resources and tasks. An edge arises in the network $(v_i, v_j) \in E$ if two nodes exhibit interaction behaviour. By assigning weights to edges, which are computed depending on techniques including frequency of behaviour, time decay or semantic similarity, one may also express the strength of social behaviours, so improving the dynamic representation of the network structure.

Furthermore, the multimodal feature variable h_i of every node will be fed into the model as an initial representation for feature propagation and updating to match the input criteria of the GNN model. Apart from reflecting the existence and direction of social interactions, the existence of edges regulates the path and extent of information spread in the graph (Yi et al., 2020). For learning route optimisation, for instance, edges created by regular interactions between students would have more impact and assist to identify possible cooperative learning groups or patterns in resource migration.

Furthermore, reflecting heterogeneity, that is, the existence of several kinds of nodes with various edge relationships is the graph structure. Under this framework, the interaction logic between several node kinds is distinct; for example, user-resource is generally expressed as an access relationship while resource-resource may imply a content suggestion chain. Thus, the graph modelling process should clarify node types and edge types to build the basis for the later inclusion of heterogeneous GNN.

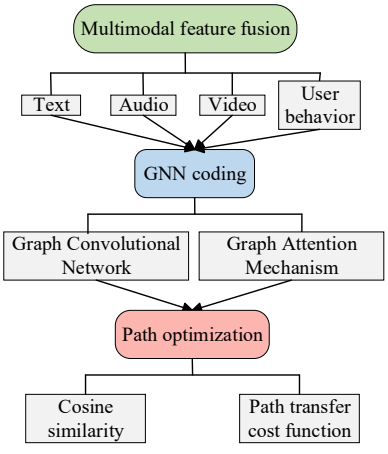
Constructing succinct but structurally rich social network graphs helps one to efficiently combine individual behaviours and social interactions in multimodal social data. This allows GNN to better model the higher-order links between nodes, therefore offering structured support and reasoning capacity for the optimisation of English learning routes.

3 Design of path optimisation model based on GNN

3.1 Model architecture and path optimisation mechanism

This work proposes an MMP-GENLPO model to construct a GNN architecture combining multi-source characteristics based on multimodal social network data to attain intelligent optimisation of English learning paths. See Figure 1 to show the three basic components of the model: the multimodal feature fusion module, the GNN encoding module, and the route optimisation mechanism, which together enable the accurate inference and dynamic suggestion of the learning path.

Figure 1 Design of the MMP-GENLPO model (see online version for colours)



3.1.1 Multimodal feature fusion module

Key component of the front-end of the model, the multimodal feature fusion module seeks to naturally integrate heterogeneous information from many sources, including text, audio, video, and user behaviours, so forming a unified and expressive representation of node features, so providing a strong data basis for the subsequent GNN-based learning path optimisation. Unlike the definition and preliminary study of multimodal data in Chapter 2, this module concentrates on how to efficiently integrate the features of each modality at the model level, improve the quality of the fused representation, and so enhance the model's understanding and prediction of English learning behaviours in complex social environments.

The first step of fusion is specifically feature extraction of every modality. Text data is encoded by BERT, a pre-trained language model based on the transformer architecture, which is capable of capturing contextual semantic relationships in depth and effectively extracting rich semantic representations at the sentence level (Acheampong et al., 2021); audio data is transformed into spectrogram form, and convolutional neural network (CNN) are utilised to capture fine-grained features such as rhythm, tempo, and emotional colours of the speech; and video modality is encoded by visual transformer to encode dynamic video frame sequences and analyse the learner's facial expressions and body movements, which are visual information reflecting the learner's attention and emotional

state; and the behavioural data modality uses a multilayer perceptron (MLP) to encode the user's behavioural statistics such as liking, commenting, sharing, etc. to reveal the learner's activity and interest preferences in the social network (Deldjoo et al., 2020). These specialist encoders transfer the feature variables $z_i^{(m)}$ produced for every modality into a consistent feature space, therefore guaranteeing that variables from many modalities are compatible and comparable inside the same dimension d. The feature variables so transfer into a homogenous feature space:

$$\alpha_i^{(m)} = \frac{\exp\left(q^\top \tanh\left(Wz_i^{(m)} + b\right)\right)}{\sum_{k=1}^{M} \exp\left(q^\top \tanh\left(Wz_i^{(k)} + b\right)\right)} \tag{3}$$

where the nonlinear transformation tanh enhances the nonlinear fitting capacity of the model representation; W, b, and q are learnt during training. The approach lets the fusion process be adaptive, automatically spotting and stressing the modal characteristics that most help the learning path to be optimised, hence, reducing noise and redundant input. Synthesised as a weighted summation, the fused node feature variables:

$$h_i = \sum_{m=1}^{M} \alpha_i^{(m)} z_i^{(m)} \tag{4}$$

The fusion module presents residual linkage, which sums the original input features x_i with the fusion variables, therefore improving the stability and expressiveness of model training:

$$h_i^{\text{res}} = h_i + x_i \tag{5}$$

By allowing information to be freely transferred between layers and hence, preventing too significant feature loss, the design essentially solves the gradient vanishing issue in deep network training. Furthermore, layer normalisation is applied to normalise the features following residuals, therefore lowering the scale variation between features and enhancing generalisation performance by means of faster convergence of the model.

By means of the above multimodal fusion process, the node highlights the function of important modes with dynamic weight adjustment, so significantly improving the richness and discriminative power of the feature representation. The node features not only cover the information in multiple dimensions of speech, voice, vision and behaviour.

3.1.2 GNN coding module

The main focus of this work is on the GNN coding module, which performs the important chore of naturally merging the fused multimodal node features with the complicated social network structure. Although the multimodal feature fusion module has already converted heterogeneous data including text, audio, video and user behaviour into unified node representations, these representations still lack a thorough knowledge of the relationships between nodes. Students in social networks not only exhibit different individual behaviours but also have rich interaction and impact links with one another, therefore forming a high-dimensional and complex graph structure. Therefore, it is essential to introduce a GNN, which enables the node features to dynamically reflect their positions and relationship patterns in the network by passing and aggregating the

information of the neighbouring nodes in a layer-by-layer manner. It is difficult to precisely capture the learners' behavioural dynamics and social influence by depending just on the stationary properties of the nodes. Node embedding will thus combine social structure knowledge and individual multimodal behavioural traits, hence, forming a strong basis for learning route optimisation.

This module specifically uses the fused multimodal feature variable $h_i^{(0)}$ as the initial representation of each node v_i, which is iteratively updated by means of social network node set of neighbours $N(i)$. The l-layer updating mechanism of node characteristics follows based on the fundamental concept of graph convolutional network (GCN):

$$h_i^{(l)} = \sigma\left(W^{(l)} \cdot \frac{1}{\left|N(i)\right|+1} \sum_{j \in N(i) \cup \{i\}} h_j^{(l-1)} \right) \tag{6}$$

where σ is the ReLU activation function and $W^{(l)}$ is the weight matrix of layer l. This equation shows how nonlinearly mapped by linear transformation through the average of the surrounding nodes and their own features to get the fusion of the neighbourhood information. The local network structure and the interactions among nodes can be sufficiently captured using this approach.

This module presents the graph attention mechanism (GAT), since simple average aggregation neglects the variations in the relevance of various neighbours to a node (Sun et al., 2023). Through adaptive weight computation of neighbour nodes, this approach increases the expressive capability of feature aggregation. Calculated as is the attention weight α_{ij} of neighbour node j to node i:

$$\alpha_{ij} = \frac{\exp\left(\text{LeakyReLU}\left(a^{\top}\left[Wh_i^{(l-1)} \middle\| Wh_j^{(l-1)}\right]\right)\right)}{\sum_{k \in N(i)} \exp\left(\text{LeakyReLU}\left(a^{\top}\left[Wh_i^{(l-1)} \middle\| Wh_k^{(l-1)}\right]\right)\right)} \tag{7}$$

where a is a learnable attention variable; || represents variable splicing; LeakyReLU offers a nonlinear transformation to guarantee the flexibility of weight distribution. The computed weight α_{ij} shows the significance of the neighbour node to the target node; so, the more weight is, the more neighbour information helps to update the node feature.

Together with the attention weights, the node feature update formula is changed to take weighted sum form:

$$h_i^{(l)} = \sigma\left(\sum_{j \in N(i)} \alpha_{ij} W h_j^{(l-1)} \right) \tag{8}$$

By concentrating on neighbours who have more impact on the nodes, this updating procedure helps the model to enhance the discriminative capacity and expressive richness of node embedding. Further improving the resilience and efficiency of the model is the computation of several independent attention heads in parallel using the multi-head attention mechanism to capture varied neighbourhood features (Reza et al., 2022).

Following several layers of iterations, the node representation forms a high-dimensional variable fit for the next English learning path optimisation job by combining the multimodal properties of the node itself and the structural information of its multi-order neighbours. Apart from reflecting the learner's multimodal behavioural

pattern, the variable also shows his impact and interaction structure in the social network, therefore supporting individualised path recommendation and dynamic optimisation.

By means of graph convolution and attention mechanism, the GNN coding module achieves efficient coding of multimodal node features in complex social networks, so greatly enhancing the expressive capacity of learners' features and providing strong theoretical and technical support for English learning path optimisation.

3.1.3 Path optimisation mechanism

Key module dynamically created for individual learning needs and social interactions based on multimodal node characteristics and social network structure information generated by the GNN coding module is the path optimisation mechanism. Its main objective is to intelligently plan the suggested sequence of learning resources, so enabling learners to acquire knowledge and systematically improve their language skills simultaneously and at the same time fully use the benefits of cooperative learning in social networks.

First, the model in the path optimisation process must show the link among the nodes in the path to guarantee the coherence and logic of the sequence of learning resources. The continuity of the content between the adjacent nodes in the path may be rather evaluated by computing their feature similarity. Assuming the path comprises of a sequence of nodes P, the corresponding multimodal fusion feature variable of each node is h_i, and the cosine similarity of neighbouring nodes is defined as:

$$S\left(v_i, v_{i+1}\right) = \frac{h_i \cdot h_{i+1}}{\left\|h_i\right\|\left\|h_{i+1}\right\|} \tag{9}$$

The similarity indicates the proximity of the learning materials in terms of content and presentation; a higher similarity signifies that the learning path is more coherent, which allows learners to construct a clear cognitive framework in the knowledge system (Shi et al., 2020). Based on this, the model changes the path structure and gives significant importance to linking nodes with great similarity, hence, lessening the cognitive load and reducing jump learning.

Path optimisation, meanwhile, has to include path length and learning cost as well. Too long or too complicated paths not only extend the learning time but also could cause cognitive tiredness and declining learning interest. Thus, the path transfer cost function $C(P)$ is introduced to depict the whole cost of switching between nodes in a path:

$$C(P) = \sum_{i=1}^{n-1} d\left(v_i, v_{i+1}\right) \tag{10}$$

where $d(v_i, v_{i+1})$ gauges the switching cost from node v_i to node v_{i+1} encompassing elements of changes in the difficulty of learning content, variations in resource types, and time or energy consumption. Reducing this cost helps the path optimisation process to avoid producing overly long or frequent jumping learning sequences, hence, enhancing the utility and learning efficiency of the path.

Furthermore, fully considers the impact of interactions in social networks in path optimising process. Through social interactions, students not only pick knowledge from personal resources but also get support and inspiration. The model integrates the learning status and behavioural traits of the surrounding nodes into the path planning using the

social ties acquired by the GNN, hence, improving the social fitness of the path recommendation. For instance, the advice includes materials utilised by students in the social circle with frequent interactions, therefore encouraging group learning and knowledge exchange (Sweet et al., 2020).

Mostly, the technical instruments to reach path optimisation are RL and heuristic search. The intelligent body dynamically chooses the next resources depending on the present node attributes and historical route states; the RL framework treats the learning path as a sequence of strategies. By use of constant observation of learning feedback and reward signals, the intelligent body modifies its approach to optimise the total use of the way. Conversely, heuristic search techniques consider search efficiency and result quality and mix similarity with transfer cost to rapidly filter high-quality pathways.

The capacity of the path optimisation mechanism to react to learners' behavioural changes and social network structure updates in real time reflects its dynamics, therefore facilitating live adjustment and customisation of paths. In this sense, the model not only fits various learning phases and changes in interests but also catches developing learning resources and social interactions, guaranteeing the timeliness and relevancy of recommendations.

By means of combined use of multimodal feature expression, inter-node similarity and transfer cost, and social network influence, the path optimisation mechanism enables scientific planning and dynamic optimisation of English learning paths. The technique greatly enhances the rationality of learning path and the capacity of continual development of learning effect. It also offers strong technical support for tailored learning resource recommendation in multimodal social network environment.

3.2 Modal fusion and learning strategies

Learners not only express their opinions through text communication but also rely mostly on voice dialogues, video learning and interactive activities in the process of English learning, which are entwined with each other to form a multidimensional, rich and dynamic record of learning activities. Social network platforms reflect this diversified development of these aspects. The multimodal social data we investigate in this work mostly come from English learning communities, online discussion groups and multimedia teaching platforms, and are gathered from textual data of learning posts, voice interaction records, video teaching clips, as well as the behavioural trajectories of users's liking, commenting, sharing, etc. The data are gathered concerning kind, timing, and quantity. These data show very significant variability in terms of amount, kindness and temporality, so offering a rich and thorough basis for in-depth investigation of learning processes.

This work develops a multimodal feature fusion method based on weighted summation, which maps the feature variables of every modality into a uniform feature space and merges them weighted by weighting coefficients:

$$h = \sum_{m=1}^{M} w_m x^{(m)} \tag{11}$$

where M is the number of modes; the m^{th} modal feature is $x(m)$; the model learns the weights w_m automatically to evaluate the significance of every modality to the composite feature. By means of the adaptive modification of weights, this approach dynamically

balances the information contribution of every modality, therefore preventing the expression bias resulting from inadequate information or noise interference of one modality (Huang et al., 2023). Furthermore, the fusion method considers the complementing character of multimodal information and enhances the expressiveness of important semantic and behavioural aspects, thereby enabling the in-depth knowledge of learning paths by downstream GNN.

Moreover, considering the time-varying character of user behaviours and content expressions in social network environments, this paper also presents temporal context features, so combining historical interaction records with dynamic change trends to improve the temporal sensing capability of node features. This allows the model to not only reflect the learner's interest evolution and behavioural patterns across time, therefore meeting the demand for tailored learning path optimisation, but also capture the current learning state.

This work presents a well-structured and vividly expressed node feature system by methodically evaluating and integrating multi-source heterogeneous multimodal social data, so laying a strong data basis for GNN-based English learning path optimisation. For complicated social learning situations, the approach efficiently combines textual, visual, auditory, and behavioural input, so enhancing the adaptability and prediction accuracy of the model.

4 Experiments and analyses

4.1 Datasets and assessment indicators

This work builds a multimodal social network English learning dataset to confirm the validity of the MMP-GENLPO model. Derived from a massive social network for English learning, gathered from 2023 to 2024, the dataset includes multimodal learner interaction data (this post, voice communication, learning videos and user behaviour data). Most of the users in the dataset are non-native English learners with rich social contacts and multimodal learning resource consumption patterns.

Table 1 Content of the dataset

Data type	*Description*	*Volume*	*Notes*
Text	Forum posts, comments, messages	Approximately 5,000 entries	Covers various discussion topics and feedback
Audio	Voice Q&A, spoken practice	Approximately 2,000 entries	Includes multiple speaking exercises and interactions
Video	English learning short videos	Approximately 1,000 entries	Contains listening, speaking tutorials, and demonstration videos
Behaviour	Likes, comments, shares, learning duration	Approximately 10,000 entries	Records user interactions and engagement metrics

The text data covers learning discussions, questions and answers, and feedback; the voice data includes learners's pronunciation practice and voice communication; the video data is mostly the short English learning videos and course clips shared; and the user behaviour data records social activities such likes, comments, retweets, and learning hours. These multimodal data together with social interactions among users provide a

dynamic and complex multimodal social network environment that offers a rich information basis for GNN-based learning route optimisation. Table 1 details the dataset information.

Four evaluation metrics are developed in this paper to measure learning efficiency, social interaction, stability of learning participation, and the effect of multimodal feature fusion, respectively, so enabling a comprehensive evaluation of the MMP-GENLPO model in the task of optimising English learning paths in multimodal social networks.

The optimising effect of the model-generated learning routes in terms of time and resource usage is evaluated using the normalised learning path efficiency (NLPE) metric. By means of a comparison between the entire learning time of the model-recommended path and the ideal shortest path time, NLPE specifically shows the rationality of the path design and the enhancement of learning efficiency (Liao et al., 2022). The model-recommended path is closer to the ideal one the closer the value is to 1; this will enable students effectively master knowledge in limited time, so avoiding repetitive material and inefficient learning time. Evaluating the application value of the path optimisation model in actual educational situations depends primarily on this indicator.

$$NLPE = 1 - \frac{T_{model} - T_{optimal}}{T_{\max} - T_{optimal}} \tag{12}$$

where $T_{optimal}$ is theoretical shortest learning time; T_{model} shows the entire learning time of the path produced by the model; $T_{\max}$ is the defined maximum tolerated time.

The active degree of social engagement set off by the learning process is gauged using normalised social interaction score (NSIS.). English learning is a highly interactive cognitive and communication activity; hence, the frequency and quality of social contacts directly affect the learning effects. Combining several social behaviours including the number of comments, likes, and voice exchanges on the pathways, NSIS is normalised to indicate the impact of the design of the learning paths on the promotion of social activeness. Reflecting the features of multimodal social network learning, this statistic serves to expose the capacity of the model to foster contact and cooperation among students:

$$NSIS = \frac{S_{model} - S_{\min}}{S_{\max} - S_{\min}} \tag{13}$$

where S_{model} is the social interaction score of the suggested paths; $S_{\min}$ and $S_{\max}$ are the least and maximum values of social interaction in the dataset correspondingly.

Normalised engagement consistency (NEC) gauges how steadily students participate during the path of instruction. By means of data on learning behaviours, that is, variations in the number of consecutive log-in days and the length of each learning session, NEC computes the standard deviation of engagement and normalises it to a stability indicator. Higher NEC values show that the design of the learning path can efficiently preserve learners' motivation and interest in learning, therefore lowering the phenomena of disconnection and abandonment during education:

$$NEC = 1 - \frac{\sigma_{engagement}}{\sigma_{\max}} \tag{14}$$

where $\sigma_{\max}$ is the maximum value of the engagement fluctuation; $\sigma_{engagement}$ is the standard deviation of the learning engagement matching with the model routes.

The effect of multimodal feature fusion to the general model performance is evaluated using normalised modal fusion effectiveness (NMFE). One of the main approaches of this work is multimodal data fusion; this measure is normalised to evaluate the increase of the fusion effect by means of performance difference between the fused multimodal features and the performance difference between employing just single modal characteristics. Higher NMFE values show that multimodal fusion makes good use of the complementary information between several modalities, hence, enhancing the capacity of the model to forecast and optimise challenging learning routes:

$$NMFE = \frac{P_{fusion} - P_{single}}{P_{\max} - P_{single}} \tag{15}$$

where P_{single} is the performance index in a single modality; $P_{\max}$ is the highest feasible theoretical performance; P_{fusion} is the model performance index following fusing multimodal information.

4.2 Experimental results and comparative analysis

Aiming at investigating the performance of the model and the roles of the important components in a real data environment, two representative experiments are designed in this paper to comprehensively verify the effectiveness of the proposed MMP-GENLPO model in the task of English learning path optimisation in multimodal social networks by respectively analysing the multimodal fusion capability and path optimisation mechanism.

The first experiment intends to assess the influence of the multimodal feature fusion module on English learning route optimisation effect in the MMP-GENLPO model. Given the complexity and variability of multimodal social data in real-world applications, the experiment especially plans a modal combination comparison scheme to progressively add several kinds of information sources to quantify the degree of influence of different types of modalities and their fusion mechanisms on the final path generating performance.

Four primary information modalities covering multimodal modelling's practicality and task suitability helps to improve its applicability. The platform offers a good basis for graph structure modelling and multimodal aggregation since it clearly has social aspects, highly scattered distribution of learning materials, and visible social route dependence of user learning behaviour.

This work regards each user or learning content as a node in the graph in the data preparation stage, and the attention relationship between users, interaction records, and common learning content as edges for graph development. Four groups of models are built up for comparative studies to evaluate how much the multimodal fusion technique enhances the effect of the final learning route generating, as Table 2 shows.

Whereas the GNN coding structure is constant, all comparison models add new modal combinations appropriately. Except for MMP-GENLPO, the other models only perform simple splicing or average pooling; MMP-GENLPO introduces the modal attention mechanism in the fusion process, which learns the feature representations of

different modalities in a weighted manner, so generating node representations with more discriminative and semantic depth.

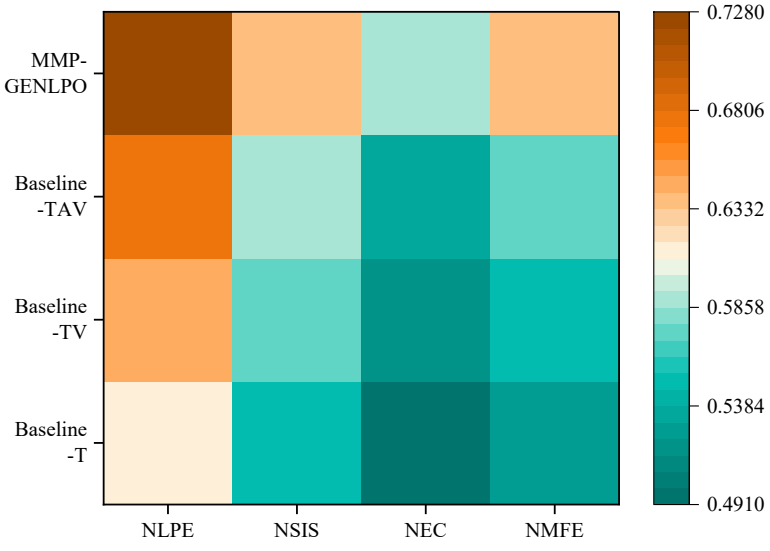
Table 2 Information on the four groups of comparison models

Model name	*Input modalities*	*Fusion applied*	*Path optimisation*
Baseline-T	Text	No	No
Baseline-TV	Text + video	No	No
Baseline-TAV	Text + video + audio	No	No
MMP-GENLPO	Text + video + audio + behaviour	Yes	Yes

The same parameter configuration is applied for all models in the training setup: Adam is the optimiser; the learning rate is set to 0.001; the batch size is 64; the maximum number of training rounds is 200; an early-stop strategy is employed to prevent overfitting. All experiments are done for five rounds and averaged to guarantee the stability and repeatability of the outcomes; the training dataset is split into training, validation, and test sets in the ratios of 80%, 10% and 10%.

All four of the evaluation measures have been normalised to the range [0, 1], with higher values denoting better performance. Figure 2 displays the experimental outcomes.

Figure 2 Experimental results on the effectiveness of multimodal feature fusion (see online version for colours)



The four groups of models show notable variations in all four indicators, according to the results. First in terms of NLPE measurements, the route efficiency of the models keeps rising when input modalities are further enriched: Baseline-T score is 0.612, Baseline-TV is 0.643, Baseline-TAV is 0.674, and the whole model MMP-GENLPO reaches the highest 0.728. While the modal alignment and fusion technique proposed in MMP-GENLPO may essentially enhance the fit between recommended paths and learning objectives, this trend illustrates that the expansion of modal information can successfully improve the fit between recommended paths and learning objectives. This

trend shows that the growth of modal information can efficiently improve the match between the recommended paths and the learning objectives; moreover, the modal alignment and fusion technique presented in MMP-GENLPO strengthens the accuracy and practicability of path building.

The four groups of models clearly differ in the NSIS dimension as well. Whereas for Baseline-TV and Baseline-TAV the index rises to 0.572 and 0.593, respectively, suggesting that the multimodal content helps to capture the possible connection between user preferences and resource semantics, the Baseline-T model depends just on textual semantics and shows low semantic interaction consistency with an NSIS score of 0.548. semantics for resources. By means of the attention mechanism, MMP-GENLPO mines the complementary relationship between modalities and increases the NSIS value (0.641), so verifying the efficiency of the fusion technique for improving semantic hierarchy interaction modelling.

The consistency of users' behaviour under the system recommendation is measured by NEC; the scores of Baseline-T and Baseline-TV are 0.491 and 0.516 respectively, which are at the basic level; the score of Baseline-TAV increases somewhat (0.538), which indicates that the audio modality contributes to stabilising users's behaviours to a certain extent. MMP-GENLPO performs better (0.587), meaning that its multimodal modelling and path optimisation method may more successfully assist users to participate in the learning process in a continuous and intentional manner, therefore preserving a high degree of consistency.

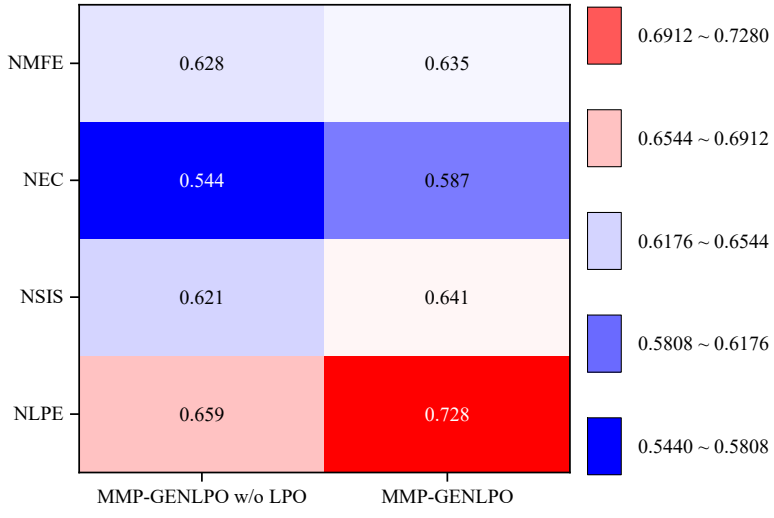
MMP-GENLPO likewise exhibits the strongest performance (0.635), which is notably better than Baseline-T (0.527), Baseline-TV (0.551) and Baseline-TAV (0.577) on the last parameter, NMFE. This result implies that while fusion strategies and structured modelling approaches can improve the representation of the feature space to more effectively support path optimisation tasks, representations built by a single or simple combination of modalities are difficult fully to capture the fine-grained features of learning resources.

Further research of the performance capabilities of MMP-GENLPO in the actual learning path generation process is therefore still necessary since the present studies have not yet identified the role of route optimisation mechanism after fusing of features. Thus, experiment 2 will concentrate on the path optimisation mechanism itself and evaluate its contribution to stability, effectiveness and personalised expression of path planning by building several sets of learning objectives and intervention strategies based on actual user behavioural trajectories. This experiment validates the practical relevance of the MMP-GENLPO model for dynamic path optimisation in complicated social settings and helps to deepen knowledge of its fundamental capabilities. Two model configurations were intended to be controlled by experiment 2:

- MMP-GENLPO (complete model) with GNN path optimisation method and multimodal feature fusion.
- MMP-GENLPO w/o LPO eliminates the path optimisation method and generates static heuristic paths in the graph using just the features retrieved during multimodal fusion.

Experiment 2 shows its experimental results in Figure 3 and still employs the four metrics to assess the path quality.

Figure 3 The role of path optimisation in English learning path generation (see online version for colours)



From the experimental results, it is evident that the full model MMP-GENLPO outperforms the comparison model MMP-GENLPO w/o LPO without introducing the path optimisation mechanism in all the four-normalisation metrics, hence, indicating that the path optimisation mechanism has a major contributing effect on improving the recommendation quality of the multimodal social English learning system.

First, on NLPE, the whole model gains 10.5%, indicating that the path optimisation mechanism can efficiently compress the length of the learning path and assist the user to attain the learning goal more rapidly without compromising the integrity of the information. This implies that the model not only emphasises the significance of individual nodes but also organises resource configurations from the whole viewpoint of the path, therefore enhancing the teaching effectiveness of the system.

Second, NSIS increases by roughly 3.2%, indicating that in resource sequencing the path optimisation strategy can improve the semantic coherence between adjacent nodes. Stated differently, students will not experience issues including abrupt topic changes or difficulty leaps when following the advised paths, so facilitating the development of a steady semantic construction process.

After the introduction of path optimisation, the user's behaviour on the learning path shows stronger consistency and concentration, the fluctuation of data such as click sequence and dwell time is lowered, and the user is more inclined to follow the recommended path to learn in a continuous manner. The whole model shows an improvement of almost 8% over the baseline on NEC. This suggests some degree of alignment between the optimal path and users' cognitive rhythm, therefore enhancing their learning initiative.

In high-dimensional modal fusion situations, the NMFE is still significant even if it has just raised by around 1.1%. The optimisation module improves the consideration of modal complementarity in the path construction process, so ensuring that the combination of several modal resources in the path is more in line with the learning task requirements.

This indicator mostly reflects the modal matching and task adaptability of the selected resources in the path.

Further from the user feedback samples, it can be observed that users are generally satisfied with the logic of the recommended content and task matching after the introduction of the path optimisation mechanism; the general difficulty curve of the path is smoother and less likely to have learning bottlenecks or repeated content. From a pragmatic standpoint, this also confirms the good worth of the process in enhancing the effect of tailored recommendations.

5 Conclusions

This work focuses on the problem of personalised optimisation of English learning paths in multimodal social networks and presents a GNN-based fusion model MMP-GENLPO, which makes complete use of multimodal data to improve the effectiveness of the recommender system in real learning scenarios. By means of modular design, the three fundamental components of the model efficiently combine the social association link between users and the complementary information between modalities, so improving the coherence, adaptability, and teaching effectiveness of the learning path.

This work intends two stages of empirical study in the experimental section. By using a multi-group ablation comparison, the first experiment confirms the beneficial effect of the multimodal fusion module in improving resource comprehension and learning path quality. The second experiment aims to evaluate the improving influence of the path optimising mechanism on structural rationality and behavioural consistency of learning paths. The MMP-GENLPO model clearly benefits in terms of path efficiency, semantic coherence, and multimodal expressiveness based on the results of the four indicators (NLPE, NSIS, NEC, NMFE), so indicating great practicality and extensibility of the model suggested in this paper in real multimodal social learning environments.

This paper still suffers some flaws notwithstanding the findings of the research. Although the self-constructed dataset simulates a real scenario, there are still limits in the size and diversity of the samples, which may influence the generalisation ability of the model in a wider English learning population. First, the GNN structure used mostly focuses on static graph modelling and fails to fully capture the temporal evolutionary characteristics of dynamic social behaviours; secondly, the path optimisation is mostly based on content relevance and behavioural continuity; hence, deeper user-level factors such as cognitive load and emotional state have not yet been introduced.

First, introducing a dynamic GNN model to capture the dynamic changes of users's interests and social relationships, so as to more realistically simulate the behavioural migration and path adjustment in the process of English learning; second, extending the depth of multimodal modelling, and introducing a multi-channel attention mechanism to enhance the inter-modal synergy (Nguyen-Phuoc et al., 2025); third, combining the user's cognitive level, learning style, real-time feedback, and information from individual portraits Third, combining individual portrait information, including user cognitive level, learning style and real-time feedback data to achieve a really meaningful path planning; fourth, further building a larger and more complex multimodal English learning dataset, and promoting the transformation of the model to the deployment of the actual online learning platform.

Based on GNN and multimodal fusion approaches, this work offers a creative and workable method for English learning path optimisation in social settings. This study not only broadens the field of intelligent education's application limit for graph learning and recommender systems but also offers fresh concepts and methodological support for the future customised development of multimodal social learning systems.

Declarations

All authors declare that they have no conflicts of interest.

References

Acheampong, F.A., Nunoo-Mensah, H. and Chen, W. (2021) 'Transformer models for text-based emotion detection: a review of BERT-based approaches', *Artificial Intelligence Review*, Vol. 54, No. 8, pp.5789–5829.

Albreiki, B., Habuza, T., Palakkal, N. and Zaki, N. (2024) 'Clustering-based knowledge graphs and entity-relation representation improves the detection of at risk students', *Education and Information Technologies*, Vol. 29, No. 6, pp.6791–6820.

Cai, H., Zheng, V.W. and Chang, K.C-C. (2018) 'A comprehensive survey of graph embedding: problems, techniques, and applications', *IEEE Transactions on Knowledge and Data Engineering*, Vol. 30, No. 9, pp.1616–1637.

Chen, N-S., Yin, C., Isaias, P. and Psotka, J. (2020) 'Educational big data: extracting meaning from data for smart education', pp.142–147, Taylor & Francis.

Chen, X., Zou, D., Xie, H. and Cheng, G. (2021) 'Twenty years of personalized language learning', *Educational Technology & Society*, Vol. 24, No. 1, pp.205–222.

Deldjoo, Y., Schedl, M., Cremonesi, P. and Pasi, G. (2020) 'Recommender systems leveraging multimedia content', *ACM Computing Surveys (CSUR)*, Vol. 53, No. 5, pp.1–38.

Dong, G., Tang, M., Wang, Z., Gao, J., Guo, S., Cai, L., Gutierrez, R., Campbel, B., Barnes, L.E. and Boukhechba, M. (2023) 'Graph neural networks in IoT: a survey', *ACM Transactions on Sensor Networks*, Vol. 19, No. 2, pp.1–50.

Gandhi, A., Adhvaryu, K., Poria, S., Cambria, E. and Hussain, A. (2023) 'Multimodal sentiment analysis: a systematic review of history, datasets, multimodal fusion methods, applications, challenges and future directions', *Information Fusion*, Vol. 91, pp.424–444.

Huang, Y., Peng, P., Zhao, Y., Xu, H., Geng, M. and Tian, Y. (2023) 'Hierarchical adaptive value estimation for multi-modal visual reinforcement learning', *Advances in Neural Information Processing Systems*, Vol. 36, pp.46724–46736.

Liao, L., Li, H., Shang, W. and Ma, L. (2022) 'An empirical study of the impact of hyperparameter tuning and model optimization on the performance properties of deep neural networks', *ACM Transactions on Software Engineering and Methodology (TOSEM)*, Vol. 31, No. 3, pp.1–40.

Lim, F.V., Toh, W. and Nguyen, T.T.H. (2022) 'Multimodality in the English language classroom: a systematic review of literature', *Linguistics and Education*, Vol. 69, p.101048.

Maged, S.A. and Mikhail, B.H. (2020) 'Deep reinforcement learning collision avoidance using policy gradient optimization and Q-learning', *International Journal of Computational Vision and Robotics*, Vol. 10, No. 3, pp.260–274.

Nguyen-Phuoc, L., Gaboriau, R., Delacroix, D. and Navarro, L. (2025) 'MMA-Net: a multimodal multitask network utilizing dual attention mechanisms for enhanced modality fusion and task exchange in cognitive load assessment', *Signal, Image and Video Processing*, Vol. 19, No. 9, pp.1–8.

Qi, M., Qin, J., Yang, Y., Wang, Y. and Luo, J. (2021) 'Semantics-aware spatial-temporal binaries for cross-modal video retrieval', *IEEE Transactions on Image Processing*, Vol. 30, pp.2989–3004.

Reza, S., Ferreira, M.C., Machado, J.J. and Tavares, J.M.R. (2022) 'A multi-head attention-based transformer model for traffic flow forecasting with a comparative analysis to recurrent neural networks', *Expert Systems with Applications*, Vol. 202, p.117275.

Shi, D., Wang, T., Xing, H. and Xu, H. (2020) 'A learning path recommendation model based on a multidimensional knowledge graph framework for e-learning', *Knowledge-Based Systems*, Vol. 195, p.105618.

Siam, S.I., Ahn, H., Liu, L., Alam, S., Shen, H., Cao, Z., Shroff, N., Krishnamachari, B., Srivastava, M. and Zhang, M. (2025) 'Artificial intelligence of things: a survey', *ACM Transactions on Sensor Networks*, Vol. 21, No. 1, pp.1–75.

Sun, C., Li, C., Lin, X., Zheng, T., Meng, F., Rui, X. and Wang, Z. (2023) 'Attention-based graph neural networks: a survey', *Artificial Intelligence Review*, Vol. 56, No. 2, pp.2263–2310.

Sweet, K.S., LeBlanc, J.K., Stough, L.M. and Sweany, N.W. (2020) 'Community building and knowledge sharing by individuals with disabilities using social media', *Journal of Computer Assisted Learning*, Vol. 36, No. 1, pp.1–11.

Yi, Y., Zhang, Z., Yang, L.T., Deng, X., Yi, L. and Wang, X. (2020) 'Social interaction and information diffusion in social internet of things: dynamics, cloud-edge, traceability', *IEEE Internet of things Journal*, Vol. 8, No. 4, pp.2177–2192.

Zhang, C., Yang, Z., He, X. and Deng, L. (2020) 'Multimodal intelligence: representation learning, information fusion, and applications', *IEEE Journal of Selected Topics in Signal Processing*, Vol. 14, No. 3, pp.478–493.

Zhu, X., Guo, C., Feng, H., Huang, Y., Feng, Y., Wang, X. and Wang, R. (2024) 'A review of key technologies for emotion analysis using multimodal information', *Cognitive Computation*, Vol. 16, No. 4, pp.1504–1530.