Online-Journals.org

iJIM International Journal of Interactive Mobile Technologies

iJIM | eISSN: 1865-7923 | Vol. 20 No. 16 (2026) | OPEN ACCESS

https://doi.org/10.3991/ijim.v20i16.62749

PAPER

Mobile Interaction Behavior Analysis for Optimizing Learning Engagement Prediction in Higher Education

Mengnan Li (✉)

North China Institute of Aerospace Engineering, Langfang, China

limengnan790929@163.com

ABSTRACT

Mobile terminals serve as the primary carriers of self-directed mobile learning among university students. Interaction data generated from these terminals can authentically reflect learners' dynamic engagement levels. Existing research, however, inadequately accommodates the non-uniform temporal characteristics of behavioral sequences, struggles to capture long-range dependencies, suffers from poor fusion quality of heterogeneous features, and exhibits limited model interpretability—thereby hindering practical deployment in instructional intervention contexts. To address these issues, a unified framework for behavioral analysis and engagement prediction was proposed. Relying on non-intrusive sensing techniques, multimodal learning behavior data were collected, and learning engagement was quantified through the fusion of subjective and objective evaluation metrics. Temporal behavioral topological associations were uncovered by constructing horizontal visibility graphs integrated with a graph attention network. A cross-modal attention mechanism was employed to adaptively fuse multi-source heterogeneous features, while multitask learning was introduced to enhance model generalization, enabling accurate prediction of learning engagement trajectories. Furthermore, a gradient-weighted attribution strategy was incorporated to identify key influencing factors, and a mobile intervention mechanism was designed to optimize learners' behavioral states. This study provides a technical reference for mobile learning behavior mining, learning state prediction, and the development of intelligent educational applications.

KEYWORDS

mobile interaction behaviors, multimodal sensing, horizontal visibility graph, graph attention network, cross-modal attention, learning engagement prediction; explainable intelligence

1 INTRODUCTION

With the iterative advancement of mobile internet and intelligent terminal technologies, mobile learning has progressively become integrated into higher education scenarios, emerging as a predominant form of self-directed learning for university students in fragmented time slots [1, 2]. Operational interactions, system responses,

Li, M. (2026). Mobile Interaction Behavior Analysis for Optimizing Learning Engagement Prediction in Higher Education. *International Journal of Interactive Mobile Technologies (iJIM)*, 20(16), pp. 104–117. https://doi.org/10.3991/ijim.v20i16.62749

Article submitted 2026-05-26. Revision uploaded 2026-06-26. Final acceptance 2026-06-26.

and spatiotemporal environmental data generated by learners on terminal devices can objectively characterize dynamic learning behavioral features, thereby providing a novel data dimension for quantifying and optimizing learning states [3, 4]. Learning engagement serves as a core benchmark for evaluating the effectiveness of mobile learning and assessing learning quality. Currently, mainstream assessment methods remain reliant on subjective questionnaires and manual statistics—approaches that suffer not only from issues such as data latency and significant subjective bias but also from an inability to accommodate long-span, highly dynamic mobile behavioral analysis scenarios [5, 6]. In the fields of mobile computing and human–computer interaction, leveraging on-device sensing techniques and deep learning algorithms to parse user behavioral characteristics, aiming to achieve accurate prediction and proactive intervention of learning states, has become a research focus in intelligent education [7, 8]. At the theoretical level, refining behavioral modeling frameworks tailored to non-uniform temporal data in mobile contexts can enrich the research paradigms of intelligent behavior analysis [9, 10]. At the application level, relevant technical outcomes can be deployed across various educational terminal platforms, facilitating the transformation of mobile learning services from passive resource recommendation toward active and personalized intelligent regulation [11].

Although extensive research has been conducted both domestically and internationally on mobile learning behavior analysis and learning engagement prediction, several shortcomings persist within the overall technical framework, thereby constraining both technological deployment and academic advancement in this domain. At the data level, most studies have relied on a single category of behavioral data for analysis, overlooking the intrinsic correlations among application operations, system states, and environmental contexts [12, 13]. Furthermore, data collection has often been characterized by short cycles and high intrusiveness, making it difficult to reconstruct authentic learner behaviors. In addition, comprehensive privacy protection mechanisms are generally lacking, which fails to meet the data requirements of longitudinal empirical studies [14]. At the modeling level, conventional time series models have frequently been employed to process behavioral data. Such algorithms are suited to uniform time series but cannot adequately accommodate the inherently irregular intervals and long-range nonlinearly coupled dynamics of mobile behaviors, thereby limiting the capacity to deeply uncover latent topological regularities within behavioral sequences [7, 9]. At the feature processing level, fusion methods for multi-source heterogeneous data have remained relatively simplistic, with direct feature concatenation being the dominant approach. Consequently, dynamic cross-modal associations are difficult to capture, and differentiated learning states—such as genuine concentration versus invalid dwell—cannot be effectively distinguished [10, 12]. In addition, existing prediction models exhibit weak overall interpretability, making it difficult to identify key behavioral factors influencing learning engagement. Most studies have been confined to algorithmic prediction without designing corresponding intervention strategies tailored to the operational characteristics of mobile terminals. As a result, research outcomes are difficult to translate into practical instructional settings [15].

To address the aforementioned technical limitations, a systematic framework for mobile learning behavior analysis and engagement optimization is constructed, through which multidimensional innovations are achieved. A cross-terminal, non-intrusive data collection system is established, and high-quality multimodal datasets are constructed by integrating subjective and objective evaluation metrics, thereby balancing data acquisition efficiency with user privacy protection. A topological graph structure is innovatively fused with a graph attention algorithm to resolve the challenge of modeling long-range dependencies in non-uniform temporal behavioral sequences. A novel cross-modal fusion mechanism is designed based on the Transformer

architecture, and model robustness is further enhanced by incorporating a multi-task learning strategy. Furthermore, multidimensional attribution algorithms are employed to interpret the prediction mechanism, and a mobile intervention module is deployed accordingly. Consequently, a complete technical closed loop encompassing data collection, feature modeling, state prediction, and behavioral intervention is formed, which can serve as a reference for behavioral analysis of similar user populations.

2 METHODOLOGY

2.1 Overview of the overall technical framework

A layered and progressive unified technical architecture is proposed, driven by mobile learning interaction data generated by university students in real-world scenarios. A complete closed-loop system is constructed to enable behavioral analysis and learning engagement optimization. The overall architecture is divided vertically into four functional layers, which are sequentially interconnected and functionally complementary. At the data layer, multidimensional mobile learning behaviors are collected, followed by data cleaning, sequentialization, and quantification of learning engagement labels, thereby providing high-quality data support for subsequent modeling. At the modeling layer, non-uniform temporal behavioral sequences are deeply analyzed through topological representation and graph neural networks, allowing hidden structural features and long-range dependencies within behavioral sequences to be uncovered. At the prediction layer, multi-source heterogeneous features are adaptively fused via a cross-modal attention mechanism, and model generalization is reinforced by incorporating a multi-task learning strategy, enabling dynamic rolling prediction of learning engagement trajectories. At the application layer, attribution analysis is performed on the prediction results, core behavioral factors influencing learning states are identified, and personalized optimization strategies are generated and delivered accordingly. Ultimately, full coverage of mobile learning behavior perception, feature modeling, state prediction, and intelligent intervention is achieved from a technical perspective. Figure 1 illustrates the unified technical framework for mobile interaction behavior analysis and learning engagement optimization.

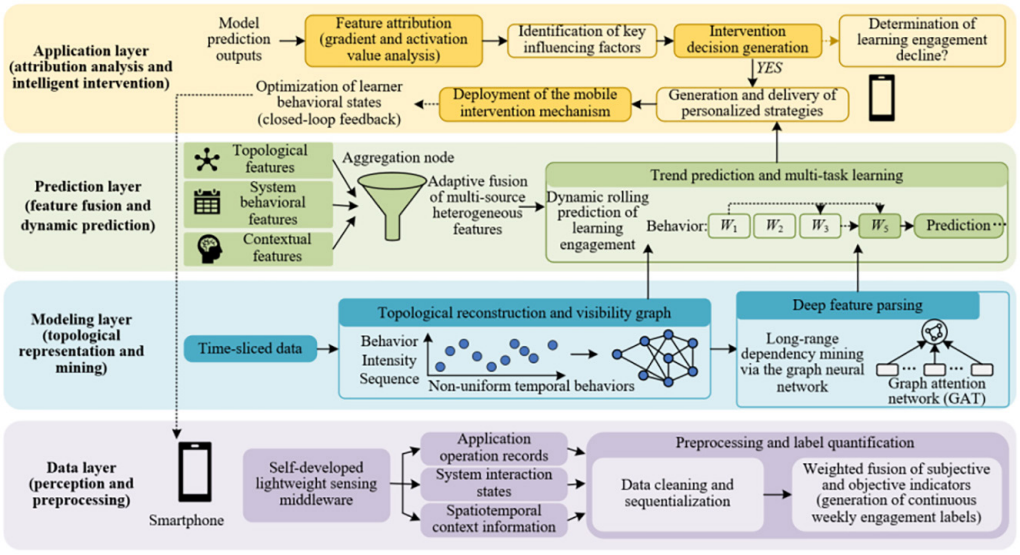

Fig. 1. Unified technical framework for mobile interaction behavior analysis and learning engagement optimization

2.2 Multimodal perception of mobile earning behaviors and data preprocessing

To accurately capture learners' dynamic interaction characteristics in authentic learning scenarios, data collection is conducted using a self-developed lightweight sensing middleware. This component is compatible with mainstream mobile operating systems and enables non-intrusive data acquisition through background silent services. To balance the acquisition quality of discrete interaction events and continuous system states, a dual strategy combining event-triggered and periodic sampling is embedded within the collection mechanism. Three categories of behavioral data—application operation records, system interaction states, and spatiotemporal context information—are continuously recorded throughout a complete academic semester. All privacy-related raw information is anonymized and desensitized locally on the device, thereby ensuring data integrity while satisfying compliance requirements for data collection. On this basis, learning engagement levels are quantified through the fusion of subjective and objective evaluation indicators. Weekly self-report scale scores and course academic performance data are standardized and then fused with weight coefficients of 0.6 and 0.4, respectively, generating continuous weekly engagement labels $y_t \in [0, 1]$, which serve as the ground truth for model training.

Raw mobile behavioral data generally suffer from noise redundancy, heterogeneous temporal granularity, and missing samples, making them unsuitable for direct model training. A multi-stage preprocessing scheme is therefore adopted to perform data cleaning and structured encoding. First, system anomaly data, short-duration accidental touch behaviors, and invalid data from non-learning periods are removed, reducing the interference of noise on modeling effectiveness at the source. Subsequently, a fixed time slice of five minutes is set, and temporal alignment and resampling are performed on multi-granularity heterogeneous data. Statistical features of various behaviors within each time slice are aggregated, and missing samples are imputed using linear interpolation or mask labeling. To meet the input requirements of subsequent models, unified behavioral feature representation is achieved through the combination of one-hot encoding and word embedding, thereby optimizing the semantic representation efficiency of high- and low-frequency behaviors. After preprocessing, the learning behaviors of a single participant within a single week are encapsulated as a complete temporal sequence $X_i = \{x_1, x_2, ..., x_L\}$, where $x_\tau \in \mathrm{R}^{d_x}$ denotes the multidimensional behavioral feature vector corresponding to the τ-th time slice, and L represents the total number of time slices within the week.

2.3 Structural modeling of behavioral sequences based on the horizontal visibility graph and graph attention network

As shown in Figure 2, the abstract modules from Figure 1 are refined into concrete algorithmic-level implementation details, with three core micro-components being illustrated in detail: the horizontal visibility graph reconstruction and visibility criterion on the left, the graph attention network-based node feature encoding and attention mechanism in the middle, and the multi-task fusion and collaborative optimization mechanism at the top. The principles of behavioral sequence structural modeling based on the horizontal visibility graph and graph attention network are introduced below.

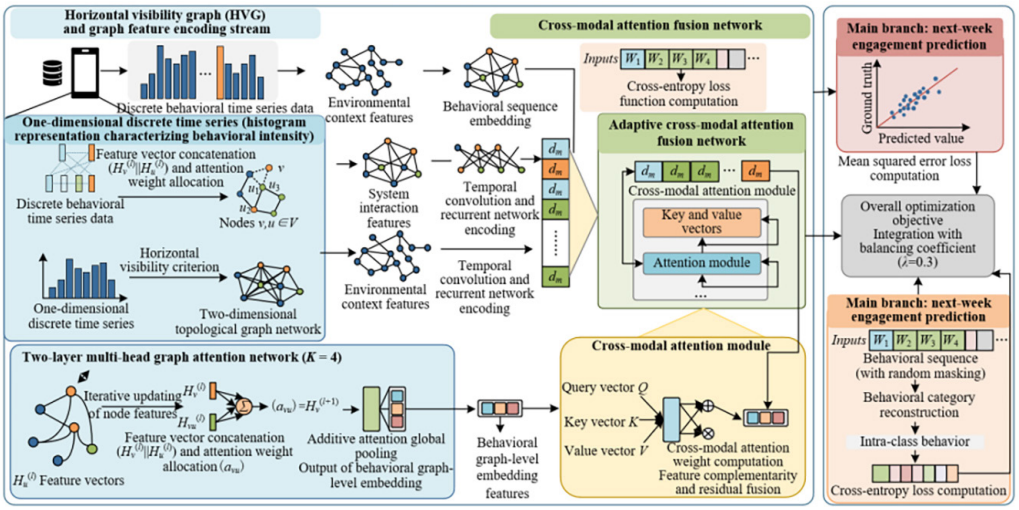

Fig. 2. Model architecture for learning engagement prediction based on the topological graph structure and cross-modal fusion

Mobile learning interaction behaviors are inherently characterized by discrete temporal distribution and irregular inter-event intervals. Conventional time series modeling approaches can only capture short-term dependencies between adjacent time steps, rendering them incapable of representing latent coupling relationships among temporally distant behaviors. To overcome the limitations of time series modeling, one-dimensional behavioral sequences are transformed into two-dimensional topological structures, and a structured reconstruction of behavioral data is accomplished through the horizontal visibility graph. The normalized behavioral intensity within each time slice is adopted as the node quantification metric, with each independent time slice being mapped to a topological node v_t. Node connectivity is then determined based on the visibility criterion, with the node connection rule defined as follows:

$$\forall c \in (a,b): y_c < \min(y_a, y_b) \Rightarrow e_{ab} = 1 \tag{1}$$

where, a and b denote the indices of any two temporal nodes, c denotes all intermediate temporal nodes within the interval between these two nodes, y_a, y_b, and y_c represent the normalized behavioral intensities of the corresponding nodes, and e_{ab} is a binary connectivity indicator. This construction mechanism disregards temporal distance constraints, enabling efficient establishment of cross-period behavioral node associations and fully recovering long-range dependency characteristics inherent in mobile behaviors. Based on this rule, a topological graph $G_t = (V, E)$ is constructed, where V is the set of all temporal nodes and E is the set of node-connecting edges. Structural metrics, including global efficiency, average clustering coefficient, and power-law exponent, are simultaneously extracted to quantify learners' overall behavioral patterns and dynamic fluctuation characteristics from a topological perspective.

To deeply mine the differentiated values of nodes within the topological network and to finely characterize the influence of key behaviors on learning states, a two-layer multi-head graph attention network is constructed for node feature encoding. The number of attention heads is set to $K = 4$, enabling multi-dimensional capture of node association features while balancing computational cost. The network computes the association strength between nodes using a learnable weight matrix and an attention vector. Dynamic weights adapted to mobile learning scenarios are obtained

through function activation and normalization, with node features being iteratively updated layer by layer. The core computational process is formalized as follows:

$$e_{vu} = LeakyReLU\left(a^{\top}[Wh_v^{(l)} \| Wh_u^{(l)}]\right) \tag{2}$$

$$\alpha_{vu} = \frac{exp(e_{vu})}{\sum_{k \in N_v \cup \{v\}} exp(e_{vk})} \tag{3}$$

$$h_v^{(l+1)} = \|_{k=1}^{K} \sigma\left(\sum_{u \in N_v \cup \{v\}} \alpha_{vu}^{(k)} W^{(k)} h_u^{(l)}\right) \tag{4}$$

where, W denotes a learnable shared weight matrix; a is an attention feature vector; $h_v^{(l)}$ and $h_u^{(l)}$ represent the feature vectors of the target node v and its neighboring node u at the l-th network layer, respectively; $\|$ denotes vector concatenation; N_v represents the set of neighboring nodes of the target node v; a_{vu} is the normalized attention weight; $\alpha_{vu}^{(k)}$ and $W^{(k)}$ denote the weight coefficient and parameter matrix corresponding to a single attention head, respectively; and σ is a non-linear activation function. During model training, additive attention global pooling is performed on all node features output by the network, yielding a graph-level embedding z_t^{graph} that characterizes the overall behavioral pattern of a single week. Additionally, an auxiliary task for classifying learning engagement patterns is introduced, with cross-entropy loss being used to constrain network parameter updates. Four typical learning engagement patterns are employed to optimize the representational direction of the embedding features, further strengthening the intrinsic association between topological features and learners' engagement states.

2.4 Cross-modal attention fusion network and learning engagement trend prediction

Mobile multi-source learning data are inherently heterogeneous, with distinct differences in representation spaces and information densities across modalities. Simple feature concatenation fails to capture dynamic cross-modal dependencies and cannot reliably distinguish complex behavioral states such as focused engagement versus invalid page dwell. In this study, three categories of data are first encoded differentially. Temporal convolutional networks are used to compress behavioral sequence features, recurrent networks are employed to encode system behavioral data, and contextual features are dimensionally expanded. All modal features are then uniformly mapped into a $\mathrm{R}^{K \times d_m}$-dimensional space, where K denotes the temporal step length and d_m represents the feature dimension of a single modality. The regularized features are subsequently fed into a six-layer Transformer encoder for deep representation learning. The network hierarchy is constructed by combining multi-head self-attention with feed-forward units, along with layer normalization and residual connections to mitigate deep training deficiencies. The basic update equations are as follows:

$$H'^{(l)} = LayerNorm\left(H^{(l-1)} + MSA(H^{(l-1)})\right) \tag{5}$$

$$H^{(l)} = LayerNorm\left(H'^{(l)} + FFN(H'^{(l)})\right) \tag{6}$$

where, $H^{(l-1)}$ and $H^{(l)}$ denote the input and output features of the l-th encoder layer, respectively, $H'^{(l)}$ is the intermediate feature after processing by the self-attention

module, and *MSA* and *FFN* represent the multi-head self-attention unit and the feed-forward network unit, respectively. To facilitate deep interaction among heterogeneous information, a cross-modal attention module is embedded within the intermediate layers of the encoder. Behavioral sequence features are adopted as the primary query, while system behavior and environmental context features are integrated to construct key and value vectors. Attention weights are adaptively allocated according to the contribution of modal information under different contextual scenarios, enabling precise interpretation of the influence of external conditions on learning behaviors. The specific computation is formalized as follows:

$$Q = H_{seq}^{(l)} W_Q,\ K = H_{sys}^{(l)} \| H_{ctx}^{(l)} W_K, V = H_{sys}^{(l)} \| H_{ctx}^{(l)} W_V \tag{7}$$

$$O_{cross} = Softmax\left(\frac{QK^{\top}}{\sqrt{d_k}}\right)V \tag{8}$$

where, $H_{seq}^{(l)}$, $H_{sys}^{(l)}$, and $H_{ctx}^{(l)}$ are the hierarchical features of the behavioral sequence, system behavior, and environmental context, respectively; W_Q, W_K, and W_V are learnable projection matrices associated with the query, key, and value transformations; Q, K, and V represent the query, key, and value vectors of the attention mechanism, respectively; $\|$ denotes feature concatenation; d_k is the dimension of the query vector; and O_{cross} is the final output feature of the cross-modal attention module. The output of this module is residually fused with the original behavioral features, enabling bidirectional complementarity of multimodal information and fundamentally enhancing the model's capacity to discriminate between differentiated learning states.

Longitudinally collected mobile behavioral data generally suffer from sparse labeling and strong temporal fluctuations. A single regression task can easily lead to model overfitting and cannot adequately adapt to complex and variable mobile learning scenarios. To address this issue, a multi-task learning framework is constructed, in which the overall model performance is optimized through collaborative training of primary and auxiliary tasks. The primary task targets the prediction of learning engagement levels in the upcoming week, with the distance between continuous predicted values and ground-truth labels being quantified using mean squared error. The loss function is defined as:

$$L_{main} = \frac{1}{N}\sum_{i=1}^{N}(\hat{y}_{t+1}^{(i)} - y_{t+1}^{(i)})^2 \tag{9}$$

where, N denotes the total number of samples in a single batch, and $\hat{y}_{t+1}^{(i)}$ and $y_{t+1}^{(i)}$ represent the model-predicted value and the ground-truth label, respectively, of learning engagement for the i-th participant in the future week. For the auxiliary task, a sequence masking reconstruction strategy is introduced, in which features of a portion of time slices are randomly masked, and the corresponding behavioral categories are then reconstructed. Cross-entropy loss is employed to constrain the model in learning the intrinsic distributional regularities of behavioral sequences:

$$L_{aux} = -\frac{1}{N_{mak}}\sum_{i=1}^{N_{mak}} \log p(type_i \mid H_{maked}) \tag{10}$$

where, N_{mak} is the total number of behavioral time slices masked, and $p(type_i \mid H_{maked})$ denotes the posterior probability predicted by the model for the i-th

behavioral category given the incomplete features H_{marked}. A balancing coefficient $\lambda = 0.3$ is set to adjust the weights of the two tasks, and the two loss functions are integrated to obtain the overall optimization objective:

$$L = L_{main} + \lambda L_{aux} \tag{11}$$

where, L_{main} and L_{aux} are the loss values of the primary regression task and the auxiliary masked reconstruction task, respectively, and the hyperparameter λ is used to balance the contribution of the two tasks to parameter updates. Throughout the training process, a rolling-window training strategy is adopted. Multimodal data from four consecutive weeks are used as input to achieve temporal prediction of the learning engagement state in the fifth week. Effective samples are augmented through window sliding, aligning with the practical application mode of online dynamic prediction on mobile terminals.

3 EXPERIMENTAL RESULTS AND ANALYSIS

3.1 Experimental configuration

The experimental data collection period covered 16 complete teaching weeks, during which mobile learning interaction behaviors of multiple university students were continuously recorded. The collected data consisted of three heterogeneous modalities: application operations, system states, and spatiotemporal context. The dataset was temporally partitioned into training, validation, and test sets, with proportions of 60%, 20%, and 20%, respectively. The test set was composed of samples from the end of the temporal sequence, thereby avoiding data leakage from the outset and aligning with the application scenario of online evaluation.

Model training was performed on a high-performance deep learning server. Mobile data collection, functional testing, and intervention experiments were deployed on multiple mainstream Android and iOS terminal devices. The algorithms were implemented using a general-purpose deep learning framework and graph computing libraries. An early stopping strategy was introduced during training to mitigate overfitting. The network architecture and training hyperparameters were configured below. The overall architecture comprised a six-layer Transformer encoder, with the number of multi-head attention heads set to eight and the feed-forward network dimension configured to 2048. The graph attention network adopted a two-layer structure with four attention heads for feature learning. The unified single-modality feature dimension was 128, and the temporal step length was set to 7. The sequence masking ratio was set to 15%, the multi-task loss balancing coefficient was 0.3, and the rolling observation window length was set to four weeks. The training batch size was 32, and the maximum number of epochs was limited to 100. Parameter updates were performed using the AdamW optimizer, with an initial learning rate of 1×10^{-4} and a weight decay coefficient of 5×10^{-5}. Training was terminated when no improvement in the validation set metrics was observed for ten consecutive epochs.

3.2 Baseline model comparison experiment

The overall performance of the proposed complete framework was compared against conventional machine learning and mainstream temporal deep learning

models to validate the technical advantages of the topological structure modeling and cross-modal attention fusion strategy in mobile multimodal behavior prediction tasks. The comparison groups were divided into two major categories: machine learning and temporal deep learning, including random forest, extreme gradient boosting, long short-term memory, gated recurrent unit, temporal convolutional network, and the original Transformer without the horizontal visibility graph-graph attention network module. All models were trained with identical input data, training strategies, and base hyperparameters. The experimental results are presented in Table 1. In the table, * denotes $p < 0.05$ and ** denotes $p < 0.01$, indicating statistically significant differences compared with the proposed method.

Table 1. Comprehensive performance comparison of different models

Model	Root Mean Square Error (Mean ± Standard Deviation)	Mean Absolute Error (Mean ± Standard Deviation)	Coefficient of Determination (R^2)	Classification Accuracy (%)	Inference Latency (ms)
Random forest	0.186 ± 0.008**	0.152 ± 0.007**	0.612	72.36 ± 1.12**	12.3
Extreme gradient boosting	0.174 ± 0.007**	0.141 ± 0.006**	0.657	74.15 ± 0.98**	14.6
Long short-term memory	0.168 ± 0.006**	0.135 ± 0.005**	0.681	75.82 ± 0.86**	28.5
Gated recurrent unit	0.165 ± 0.006**	0.132 ± 0.005**	0.694	76.14 ± 0.79**	27.2
Temporal convolutional network	0.159 ± 0.005**	0.127 ± 0.004**	0.713	77.53 ± 0.82**	31.8
Original Transformer	0.147 ± 0.005**	0.118 ± 0.004**	0.746	79.26 ± 0.75**	36.4
Proposed method	0.125 ± 0.004	0.098 ± 0.003	0.829	85.71 ± 0.63	41.2

In terms of regression metrics, the conventional machine learning models exhibited relatively high overall errors and low goodness-of-fit, rendering them incapable of adequately characterizing the complex nonlinear features of mobile learning behaviors. The temporal deep learning models showed progressively improved performance, with the original Transformer, relying on its self-attention mechanism, achieving the strongest baseline. However, intergroup comparisons revealed highly significant differences between the baseline models and the proposed method. Compared with the optimal baseline model, the proposed method reduced root mean square error and mean absolute error by 14.9% and 16.9%, respectively, increased the coefficient of determination (R^2) by 0.083, and improved classification accuracy by 6.45 percentage points. Conventional temporal models can only capture short-term dependencies between adjacent time steps and cannot adequately adapt to the irregular intervals and prominent long-range influence characteristics of mobile behaviors. In contrast, the combination of the horizontal visibility graph and the graph attention network effectively uncovered cross-period topological associations among behaviors. Regarding on-device inference efficiency, the latency of the proposed model was slightly higher than that of the lightweight baseline models, but the overall latency remained within 50 ms, fully satisfying the operational requirements for real-time prediction on mobile terminals. Taken together, these results indicate that the proposed framework achieves a favorable balance between prediction accuracy and terminal deploy ability.

3.3 Core component ablation experiment

Through an ablation design involving the removal of single modules and the combined removal of multiple modules, the independent contributions and synergistic effects of horizontal visibility graph topological modeling, graph attention network embedding, and the masked reconstruction auxiliary task were quantitatively analyzed, thereby verifying the necessity of each functional module. Four comparison configurations were established, including three single-module removal experimental groups and one dual-module combined removal group. The complete model was used as the reference benchmark, and the experimental environment was kept consistent across all groups. The test results are presented in Table 2.

Table 2. Core component ablation experiment results

Experimental Configuration	Root Mean Square Error (Mean ± Standard Deviation)	Mean Absolute Error (Mean ± Standard Deviation)	Coefficient of Determination (R^2)	Classification Accuracy (%)
Removal of the horizontal visibility graph module	0.151 ± 0.006	0.122 ± 0.005	0.721	78.94 ± 0.81
Removal of the graph attention network module	0.144 ± 0.005	0.116 ± 0.004	0.745	80.27 ± 0.76
Removal of the masked reconstruction auxiliary task	0.133 ± 0.004	0.107 ± 0.004	0.783	82.19 ± 0.68
Simultaneous removal of the horizontal visibility graph and graph attention network modules	0.163 ± 0.007	0.134 ± 0.006	0.679	75.42 ± 0.93
Complete proposed model	0.125 ± 0.004	0.098 ± 0.003	0.829	85.71 ± 0.63

Comparison of the data across groups revealed that the removal of any module led to varying degrees of decline in overall model performance. When only the horizontal visibility graph module was removed, the increase in model error was the largest, indicating that the transformation from temporal sequence to topological structure is a core component for capturing long-range dependencies in mobile behaviors. The absence of this module directly resulted in the loss of long-range behavioral association features. After the removal of the graph attention network module, adaptive weighted encoding of node features within the topological network could no longer be performed, thereby weakening the representational capacity of key behaviors, with the performance degradation being the second largest. The masked reconstruction auxiliary task enabled the extraction of intrinsic distributional regularities from behavioral sequences and effectively alleviated the overfitting problem caused by sparse labeled data on mobile terminals. After the removal of this task, a modest decline in model generalization ability was observed. When both the horizontal visibility graph and graph attention network modules were removed simultaneously, model performance dropped to the lowest level among all experimental configurations, further demonstrating the strong synergistic relationship between topological modeling and graph feature encoding. These results collectively indicate that the three modules each fulfill distinct roles while cooperating with one another, together constituting a complete feature learning system, and that the proposed module combination scheme is fully justified.

3.4 Effectiveness experiment of the cross-modal attention mechanism

To validate the performance advantages of dynamic cross-modal attention over single-modality modeling and direct feature concatenation and to supplement fine-grained modality combination comparisons, the capability of the proposed model to distinguish authentic learning engagement from spurious dwell behaviors on mobile terminals was examined. Five comparison configurations were established, covering a single behavioral modality, two-modality combinations, three-modality direct concatenation, and the proposed cross-modal attention fusion scheme. The recognition accuracy for spurious engagement was introduced as a scenario-specific evaluation metric. The experimental results are presented in Table 3.

When relying solely on the single behavioral sequence modality, the model exhibited the poorest performance across all metrics, indicating that system interaction states and spatiotemporal context contain substantial information that influences learning engagement. The two-modality combination schemes showed improved performance, but the improvement was limited due to insufficient cross-modal information interaction. Three-modality direct concatenation merely performed a simple superposition at the data dimension without modeling the dynamic associations across different modalities and consequently exhibited weak ability to recognize spurious engagement behaviors. The proposed cross-modal attention mechanism adaptively allocates weights to each modality according to the real-time context, thereby establishing deep associations among behavioral operations, system states, and environmental information. In the spurious engagement recognition task, this scheme achieved an accuracy of 88.72%, which was substantially higher than that of the other comparison groups. As evidenced by the visualization of attention weights, the model can simultaneously associate two types of features—prolonged page dwell and frequent application switching—enabling accurate differentiation between focused learning and invalid operations. These results fully demonstrate the adaptability of the proposed mechanism to complex mobile interaction scenarios.

Table 3. Performance comparison of multimodal fusion schemes

Fusion Scheme	Root Mean Square Error (Mean ± Standard Deviation)	Mean Absolute Error (Mean ± Standard Deviation)	Coefficient of Determination (R^2)	Classification Accuracy (%)	Spurious Engagement Recognition Accuracy (%)
Behavioral sequence modality only	0.162 ± 0.007	0.131 ± 0.006	0.687	76.58 ± 0.88	70.25 ± 1.21
Behavioral sequence + system behavior modality	0.149 ± 0.006	0.120 ± 0.005	0.732	79.11 ± 0.79	76.33 ± 1.05
Behavioral sequence + context modality	0.146 ± 0.005	0.117 ± 0.004	0.741	80.04 ± 0.72	74.18 ± 1.13
Three-modality direct concatenation	0.140 ± 0.005	0.112 ± 0.004	0.762	80.63 ± 0.69	79.46 ± 0.97
Proposed cross-modal attention fusion	0.125 ± 0.004	0.098 ± 0.003	0.829	85.71 ± 0.63	88.72 ± 0.84

3.5 Quantitative evaluation experiment of model interpretability

The comprehensive capability of the two-dimensional joint attribution method was evaluated from multiple perspectives. The differences between single-attribution schemes and the joint scheme were compared, and the dataset was stratified according to different levels of learning engagement to perform a fine-grained analysis of attribution effectiveness. Three experts in the field of educational technology were invited to annotate the key behaviors within the samples. Three attribution schemes were tested: attention-weight-based attribution alone, gradient–activation-value-based attribution alone, and two-dimensional joint attribution. In addition to overall sample statistics, the dataset was further stratified into high, medium, and low engagement levels. Three metrics—normalized discounted cumulative gain (nDCG), mean ranking error, and Pearson correlation coefficient—were adopted for evaluation. The results are presented in Figure 3 and Table 4.

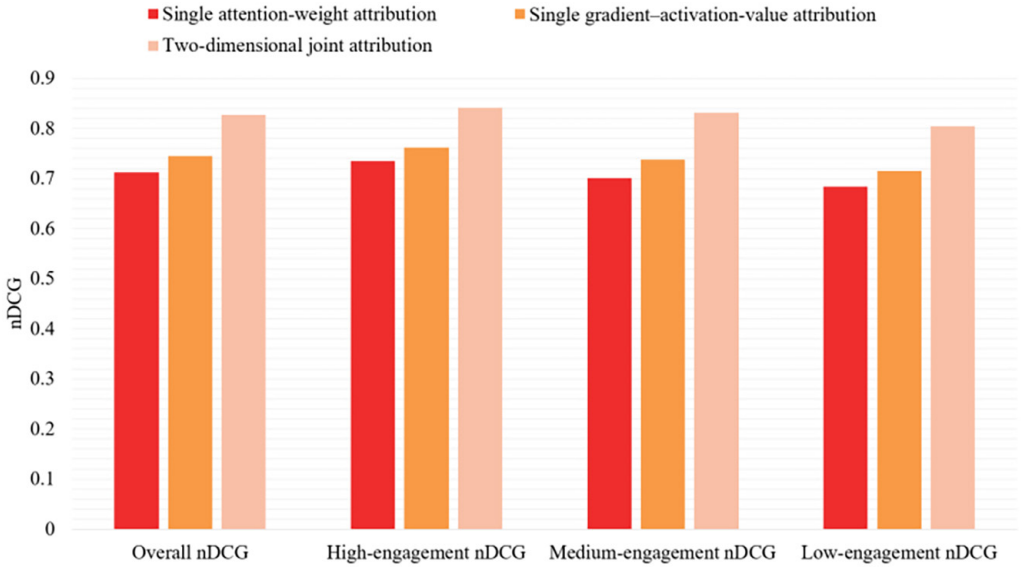

Fig. 3. nDCG comparison of different attribution schemes

Table 4. Comparison of mean ranking error and Pearson correlation coefficient across different attribution schemes

Attribution Scheme	Mean Ranking Error	Pearson Correlation Coefficient
Single attention-weight attribution	0.127	0.703
Single gradient–activation-value attribution	0.104	0.746
Two-dimensional joint attribution	0.061	0.852

The overall nDCG values of the two single-attribution schemes were both below 0.75, with relatively high mean ranking errors, indicating substantial deviation from the expert-annotated results. Attention-weight attribution excels at capturing association logic across temporal and modal dimensions but inadequately characterizes the contribution magnitude of raw features. Gradient-activation-value attribution can

precisely quantify the influence of individual feature types, yet struggles to reflect temporal dependencies among behaviors. The proposed two-dimensional joint attribution integrates the strengths of both approaches, achieving an overall nDCG of 0.827, which meets the practical application threshold, while maintaining stable performance across samples of different engagement levels. The mean ranking error was reduced to 0.061, and the Pearson correlation coefficient was increased to 0.852, demonstrating that both the ranking logic and the influence weights of the attribution results are highly consistent with expert judgments. This scheme can simultaneously identify key behavioral periods and core behavioral attributes, and the resulting conclusions can directly support the design of subsequent personalized intervention strategies.

4 CONCLUSION

Aiming to address the challenges of modeling interaction behaviors in mobile learning scenarios, insufficient prediction accuracy of learning engagement, and weak model interpretability, a complete technical framework spanning from data collection to intelligent intervention was constructed in this study. Long-term collection of multimodal behavioral data was accomplished through a self-developed cross-platform, non-intrusive sensing component, and learning engagement evaluation labels were generated by integrating subjective and objective indicators, thereby providing a reliable data foundation for model training. Topological representation of temporal behaviors was achieved by combining the horizontal visibility graph with a graph attention network, effectively overcoming the limitations of conventional time series models and fully capturing long-range nonlinear dependencies inherent in mobile interaction behaviors. A cross-modal attention module designed based on the Transformer architecture enabled adaptive fusion of heterogeneous features, and the incorporation of a multi-task learning strategy significantly enhanced model stability and predictive performance under sparse label conditions while allowing authentic learning states to be accurately distinguished from invalid page dwell behaviors. Furthermore, result attribution was performed by integrating attention weights and gradient features, followed by the deployment of a mobile intervention scheme. Empirical results demonstrated that this strategy effectively improved learners' engagement levels. The complete framework established a closed-loop intelligent application that integrates perception, modeling, prediction, and optimization and can serve as a feasible technical reference for similar research on mobile human–computer interaction behavior analysis.

5 REFERENCES

[1] R. Shkilev *et al.*, "Augmented reality in mobile learning: Enhancing interactive learning experiences," *International Journal of Interactive Mobile Technologies*, vol. 18, no. 20, pp. 4–15, 2024. https://doi.org/10.3991/ijim.v18i20.50795

[2] J. Wang, "Intelligent education based on mobile learning: Transitioning from traditional classrooms to adaptive learning environments," *International Journal of Interactive Mobile Technologies*, vol. 19, no. 11, pp. 51–65, 2025. https://doi.org/10.3991/ijim.v19i11.56057

[3] Z. Yu, L. Yu, Q. Xu, W. Xu, and P. Wu, "Effects of mobile learning technologies and social media tools on student engagement and learning outcomes of English learning," *Technology, Pedagogy and Education*, vol. 31, no. 3, pp. 381–398, 2022. https://doi.org/10.1080/1475939X.2022.2045215

[4] X. Wang, L. Hui, X. Jiang, and Y. Chen, "Online English learning engagement among digital natives: The mediating role of self-regulation," *Sustainability*, vol. 14, no. 23, p. 15661, 2022. https://doi.org/10.3390/su142315661

[5] N. Bergdahl, M. Bond, J. Sjöberg, M. Dougherty, and E. Oxley, "Unpacking student engagement in higher education learning analytics: A systematic review," *International Journal of Educational Technology in Higher Education*, vol. 21, no. 1, pp. 1–33, 2024. https://doi.org/10.1186/s41239-024-00493-y

[6] N. A. Johar, S. Na Kew, Z. Tasir, and E. Koh, "Learning analytics on student engagement to enhance students' learning performance: A systematic review," *Sustainability*, vol. 15, no. 10, p. 7849, 2023. https://doi.org/10.3390/su15107849

[7] F. Chen and Y. Cui, "Utilizing student time series behaviour in learning management systems for early prediction of course performance," *Journal of Learning Analytics*, vol. 7, no. 2, pp. 1–17, 2020. https://doi.org/10.18608/jla.2020.72.1

[8] S. B. Dias, S. J. Hadjileontiadou, J. Diniz, and L. J. Hadjileontiadis, "DeepLMS: A deep learning predictive model for supporting online learning in the Covid-19 era," *Scientific Reports*, vol. 10, no. 1, pp. 1–17, 2020. https://doi.org/10.1038/s41598-020-76740-9

[9] V. Sher, M. Hatala, and D. Gašević, "When do learners study? An analysis of the time-of-day and weekday-weekend usage patterns of learning management systems from mobile and computers in blended learning," *Journal of Learning Analytics*, vol. 9, no. 2, pp. 1–23, 2022. https://doi.org/10.18608/jla.2022.6697

[10] S. Mu, M. Cui, and X. Huang, "Multimodal data fusion in learning analytics: A systematic review," *Sensors*, vol. 20, no. 23, p. 6856, 2020. https://doi.org/10.3390/s20236856

[11] N. Valle, P. Antonenko, K. Dawson, and A. C. Huggins-Manley, "Staying on target: A systematic literature review on learner-facing learning analytics dashboards," *British Journal of Educational Technology*, vol. 52, no. 4, pp. 1724–1748, 2021. https://doi.org/10.1111/bjet.13089

[12] M. Cukurova, M. Giannakos, and R. Martinez-Maldonado, "The promise and challenges of multimodal learning analytics," *British Journal of Educational Technology*, vol. 51, no. 5, pp. 1441–1449, 2020. https://doi.org/10.1111/bjet.13015

[13] W. Chango, J. A. Lara, R. Cerezo, and C. Romero, "A review on data fusion in multimodal learning analytics and educational data mining," *WIREs Data Mining and Knowledge Discovery*, vol. 12, no. 4, 2022. https://doi.org/10.1002/widm.1458

[14] Q. Liu and M. Khalil, "Understanding privacy and data protection issues in learning analytics using a systematic review," *British Journal of Educational Technology*, vol. 54, no. 6, pp. 1715–1747, 2023. https://doi.org/10.1111/bjet.13388

[15] V. Swamy, S. Du, M. Marras, and T. Kaser, "Trusting the explainers: Teacher validation of explainable artificial intelligence for course design," in *LAK 2023: 13th International Learning Analytics and Knowledge Conference*, Arlington, TX, USA, 2023, pp. 345–356. https://doi.org/10.1145/3576050.3576147

6 AUTHOR

Mengnan Li has a Masters' degree in Management, Beijing Normal University. Now, she works as a Lecturer in the Academic Affairs Office of North China Institute of Aerospace Engineering. She has published 10 papers, including 2 core journals. She has also participated in 9 research projects and is also a Chief editor of 2 monographs (E-mail: limengnan790929@163.com).