ingegneria sismica
International Journal of Earthquake Engineering
Anno XLIII - Num. 1 - 2026

Deep Learning Driven Analysis System for Children's Natural Aesthetic Education Experience: Affective Cognition and Behavioral Influence Modeling

Chunyu Bao[1,*]

[1] College of Preschool Education, Qiongtai Normal University, Haikou, Hainan, China

SUMMARY: *Aiming at the recognition requirements of children's natural aesthetic education activity experience, a deep learning driven multimodal analysis system was proposed. The video sequence, speech segment, action trajectory, gaze area and environmental semantics were unified to describe the emotional cognition, behavioral influence and online feedback in natural aesthetic experience. The system consists of cross-modal semantic encoding, shared temporal backbone and relationship inference modules, which capture the collaborative changes between expression, language response, action rhythm and scene attributes. Based on the training evaluation of 8640 labeled interaction segments from 216 young children in 36 activity sessions, 80:20 split and 5-fold cross validation were used. Experimental results show that the accuracy of experience participation prediction reaches 92.1%, the accuracy of emotional state recognition reaches 89.7%, the score of scene semantic matching reaches 87.9%, and the average inference delay is 84 ms. Compared with the control method, the response gain of aesthetic education experience reaches 18.6%, and the F1 of behavior influence trajectory recognition reaches 90.2%, which reflects a good analysis and adjustment effect.*

KEYWORDS: *Deep learning; Multimodal computing; Emotion recognition; Behavior modeling*

1 Introduction

1.1 Research scenario and task definition

Children's natural aesthetic education activities are moving from empirical observation to data-driven fine-grained analysis. Woodland walking, plant identification, light and shadow observation, stone and leaf touch, and material composition can simultaneously generate multi-source signals such as expression, speech, movement, gaze and stay trajectories, which constitute a computable representation of the experience state. Middya et al. studied audio-video modal fusion emotion recognition and proposed a model-level joint learning method [1]. Kumar et al. studied emotional biomarker-driven multimodal recognition and proposed a prediction framework for complex people [2]. Mehta et al. studied learning participation detection and proposed 3D DenseNet self-attention network [3]. Rathod et al. studied children's emotion recognition and proposed a deep model combined with interpretable analysis [4]. Selim et al. studied the calculation of learning engagement and proposed the collaborative structure of visual network and temporal network [5]. These studies show that there is a foundation for building continuous state recognition links based on multimodal inputs. In this

*baoguer2010@163.com
https://doi.org/10.65102/is2026211
Published: 30 April, 2026

paper, children's natural aesthetic education experience is defined as a dynamic computational process consisting of environmental stimulus perception, emotional cognitive coding, behavior response generation and scene semantic coupling. The task is defined as four levels: experience participation prediction, emotional state recognition, scene semantic matching and behavior influence trajectory characterization. Under this definition, a single image judgment or a single speech judgment is not enough to describe the intensity of children's aesthetic investment in natural scenes. The system must put the multi-modal changes in the same activity segment into a unified time axis for analysis, and then complete the state aggregation through the segment-level, round-level and individual level three-level units. Then it provides input for subsequent model training, online feedback and individualized adjustment. Such task setting not only retains the scene characteristics of natural aesthetic education activities, but also ensures the integrity of data structure and label traceability required by computer modeling.

1.2 Goal of system construction

The goal of system construction needs to form a unified computing link around data organization, feature coding, inference output and feedback update, rather than dividing natural aesthetic education activities into independent observation links. Tang et al. studied the fusion of expression and speech features for emotion recognition, and proposed a fusion method for multi-modal collaborative analysis [6]. Le et al. studied multi-label and multi-modal emotion recognition and proposed a Transformer-based fusion and emotion-level representation learning structure [7]. Lian et al. reviewed the deep learning path of speech, text and face representation, which provided a systematic basis for multimodal feature selection [8]. Pan et al. studied the data sets, preprocessing and fusion methods of multimodal emotion recognition, and summarized the applicable boundaries of different processes [9]. Mocanu et al. studied cross-modal audio and video emotion recognition, and proposed a collaborative strategy combining attention mechanism and metric learning [10]. Based on the above research, this paper sets the goal of the system as four aspects: complete the unified collection and alignment of image frames, speech segments, action trajectories, gaze regions and environmental semantics. A cross-modal encoder that can describe the changes of emotional cognition is constructed. The temporal inference module oriented to behavior influence relationship was established. Form a closed loop of experience adaptation that supports online updates. In the engineering implementation of the system, it is also required that the training phase and the inference phase share a consistent data interface and label mapping rules, so that the model accuracy, response delay and deployment stability are in the same optimization framework. Under this target system, the system should not only recognize the immediate emotional changes in young children's activities, but also calculate the experience shift caused by material turnover, spatial switching, and change in interaction rhythm. Only when multi-source signals are compressed into trainable, interpretable, and returnable representation vectors, subsequent content adjustment suggestions and behavior response prediction can form verifiable outputs to support real-time deployment applications.

1.3 Technical Contributions

The technical contributions build on extensions of research into action perception, multimodal emotion recognition, and temporal inference in educational scenarios. Gupta et al. studied the learning engagement detection system based on facial cues and proposed a deep learning driven engagement recognition scheme [11]. Zhao et al. studied real-time classroom behavior recognition and proposed a lightweight detection structure combining Transformer and bidirectional pyramid network [12]. Fang et al. sorted out the path of machine learning

expression recognition in educational scenes, which provided reference for emotional label design and feature selection [13]. Trabelsi et al. studied real-time attention monitoring in the classroom and proposed a deep learning implementation framework for behavior recognition [14]. Dhara et al. studied emotion recognition based on EEG signals and proposed a fuzzy ensemble deep model [15]. On this basis, the technical contributions of this paper focus on three aspects. Firstly, a multi-modal experience analysis link for children's natural aesthetic education activities is constructed, and expression, speech, movement, gaze and environmental semantics are integrated into a unified coding space. Secondly, a joint modeling method of affective cognition and behavioral influence is designed, which uses cross-modal attention and temporal relationships to aggregate continuous state representations. Thirdly, an online feedback and experience adaptation mechanism is established, so that the system can complete content adjustment suggestions and interactive response updates according to the current recognition results. The above design puts the research focus on the level of computable representation, system deployment and quantitative verification. At the same time, a consistent interface is established between the data layer, the model layer and the service layer, so that the training sample construction, label return, inference cache and result interpretation can run in the same link. In this way, it not only retains the scene characteristics of natural aesthetic education activities, but also enhances the engineering correlation between model accuracy, reasoning efficiency and output stability, so that the research is closer to the writing requirements of technical journals for methodological, verification and reproducibility.

2 Related work

The analysis system of children's natural aesthetic education experience involves multiple computing directions such as emotion recognition, behavior understanding, cross-modal fusion and online inference. The advancement of existing research mainly focuses on three paths: one focuses on automatic emotion recognition and engagement perception in educational scenes, emphasizing deployable recognition models; One category focuses on emotion computing driven by physiological signals, emphasizing implicit state representation. The other category focuses on the multi-modal fusion mechanism, emphasizing the unified coding of vision, speech, text and temporal information. Related research has provided a model basis for the experience perception system, and also provides a technical reference for real-time analysis in natural scenes.

Moise et al. studied automatic emotion recognition in educational scenes and proposed a deep learning model based on five-channel EEG input [16]. This study maps finite channel neural signals into trainable emotional feature representations, illustrating that implicit states in educational activities can enter the computational link through lightweight physiological acquisition. Ramadan et al. studied multi-modal emotion recognition based on physiological signals and proposed a machine learning fusion scheme [17]. This method integrates multi-source physiological features into a unified discriminant framework, which provides a basis for the joint modeling of non-verbal cues. Yang et al. studied emotion recognition with multi-modal physiological signals and proposed the pulse feedforward neural network structure [18]. This study transformed the time-varying physiological response into a representation with neural dynamics characteristics, which enhanced the ability of the model to describe fine-grained emotional changes. The above studies show that emotional cognitive modeling is no longer limited to single expression or speech judgment, but gradually enters the stage of multi-source state collaborative analysis.

Guo et al. studied the task of multimodal emotion recognition and proposed a neural

network framework E-MFNN for multimodal emotion fusion [19]. The framework strengthens the unified representation by fusing the discriminative features of different modalities, which has direct reference value for emotion category discrimination in complex scenes. Geetha et al. studied the progress of multimodal emotion recognition supported by deep learning and systematically sorted out three aspects: model structure, fusion mode and application boundary [20]. This study shows that multimodal emotion computing is moving from feature concatenation to hierarchical collaboration and deep fusion. Zhang et al. studied deep learning emotion recognition of audio, visual and text modalities, and systematically reviewed the key methods in recent years [21]. This study proposes that the core of multimodal systems does not lie in a simple increase in input types, but in a stable implementation of timing alignment, semantic compensation, and cross-modal weight assignment. In order to more clearly compare the technical characteristics of existing research in emotion recognition, behavior perception and multi-modal fusion, this paper sorted out the core ideas, reference contents and application boundaries of related methods, and the results are shown in Table 1.

Table 1: Comparison of related studies

Research Direction	Representative References	Method Characteristics	Referable Content	Applicability Boundary
Educational emotion recognition	[16]	EEG-driven deep modeling	Implicit emotion representation	Relies on physiological data collection
Physiological multimodal recognition	[17][18]	Multi-source physiological fusion and time-varying encoding	Nonverbal state computation	High deployment cost in practical scenarios
General multimodal emotion fusion	[19]	Unified fusion network	Cross-modal joint discrimination	Limited use of scene semantics
Review of multimodal methods	[20][21]	Reviews of fusion strategies and systems	Basis for module design	Lacks implementation in specific scenarios
Learning engagement recognition	[22]	Collaborative visual and temporal modeling	Behavioral state prediction	Insufficient semantic coverage for natural aesthetic education

In the direction of participation behavior recognition, Shiri et al. studied student participation detection in e-learning environment and proposed the EfficientNetV2-L recognition method combined with recurrent network [22]. This study combines visual representation with temporal dependence, indicating that the participation state is not a single frame result, but a dynamic output after the accumulation of continuous interaction cues. This point also has a method significance for children's natural aesthetic education activities, because children's behavior influence in the process of observation, touch, listening and response is usually gradually revealed through a short sequence, and the detection result of a single moment is difficult to completely reflect the experience intensity.

Based on the existing research, it can be seen that the existing methods have formed mature technical routes in emotion recognition, participation perception and multi-modal fusion respectively, but the systematic realization of children's natural aesthetic education experience directly is still few. The study of physiological signals strengthens the calculation of hidden

states, the general fusion model strengthens the unified representation ability, and the study of participation recognition strengthens the analysis of temporal behavior. However, the environmental semantics, material properties, and interaction trajectories in natural scenes have not been included in the same analysis framework. Based on this research foundation, we organize natural scene images, children's speech responses, action trajectories and emotional state labels as unified inputs, and construct a dedicated analysis system through cross-modal semantic coding, behavior influence inference and online feedback closed loop, so that emotional cognition and behavior influence can be jointly expressed and continuously calculated in the same model. At the same time, the system design further introduces the object-level description of the scene and the state cache of the activity segment level, so that the correspondence between environmental stimuli and individual responses can be kept stable in the time dimension, which provides a reusable data basis for subsequent experience response gain analysis and behavior influence trajectory identification. Synthesizing the existing research, it can be seen that emotion recognition, participation perception and multi-modal fusion have formed a mature technical foundation. The physiological signal modeling strengthens the implicit state representation, the cross-modal fusion method enhances the unified representation ability, and the time series recognition framework improves the continuity of the behavior process analysis. There is still room for further refinement of the dedicated implementation corresponding to children's natural aesthetic education scenes, especially the need to incorporate scene semantics, material properties, interaction rhythm and emotional changes into the same computing link. Based on this, this paper constructs a multi-modal analysis system for natural aesthetic education activities, completes emotional cognitive coding, behavior influence inference and online feedback linkage under a unified label system, and provides a consistent data basis and deployment framework for subsequent system design and experimental verification.

3 System framework for analyzing children's natural aesthetic education experience

3.1 Multimodal Data flow and acquisition mechanism of natural aesthetic education scene

Children's natural aesthetic education activities have obvious characteristics of open scenes. Emotional changes and behavioral responses often appear simultaneously in the process of gaze, gait, speech, touch and object approach. If the system only relies on a single image or a single speech signal, it is difficult to completely describe the state of children's aesthetic investment in nature observation, material contact and space transformation. Based on this feature, the acquisition terminal is designed as a four-level structure of edge perception node, time synchronization module, quality screening module and sample storage module. The visual end is responsible for recording facial expression, hand contact position, body orientation and object interaction state. The voice end is responsible for recording pitch fluctuation, speech rate change, interjection and pause duration. The action end is responsible for recording step frequency, turning Angle, acceleration and short stop. The environmental side records plant category, material texture, color distribution, light intensity, and active node labels. The data of each terminal is denoised, sliced, resampled and aligned before entering the model, so as to transform the heterogeneous records in the open scene into trainable time series samples.

Fig. 1 takes "original collection -- time window segmentation -- modal synchronization -- quality screening -- sample storage" as the main line. Each activity segment is coded as a five-tuple index of "child number - activity round - scene node - object category - time window",

which facilitates the backtracking of the training phase according to the three dimensions of individual, scene and task. Explicit signals include verbal feedback, action completion, and object selection, while implicit signals include delay duration, gaze backswing, touch dwell, and rhythm change. After processing through this link, the output of the acquisition end is no longer discrete sensing records, but fragment-level tensor representations suitable for batch training and online inference.

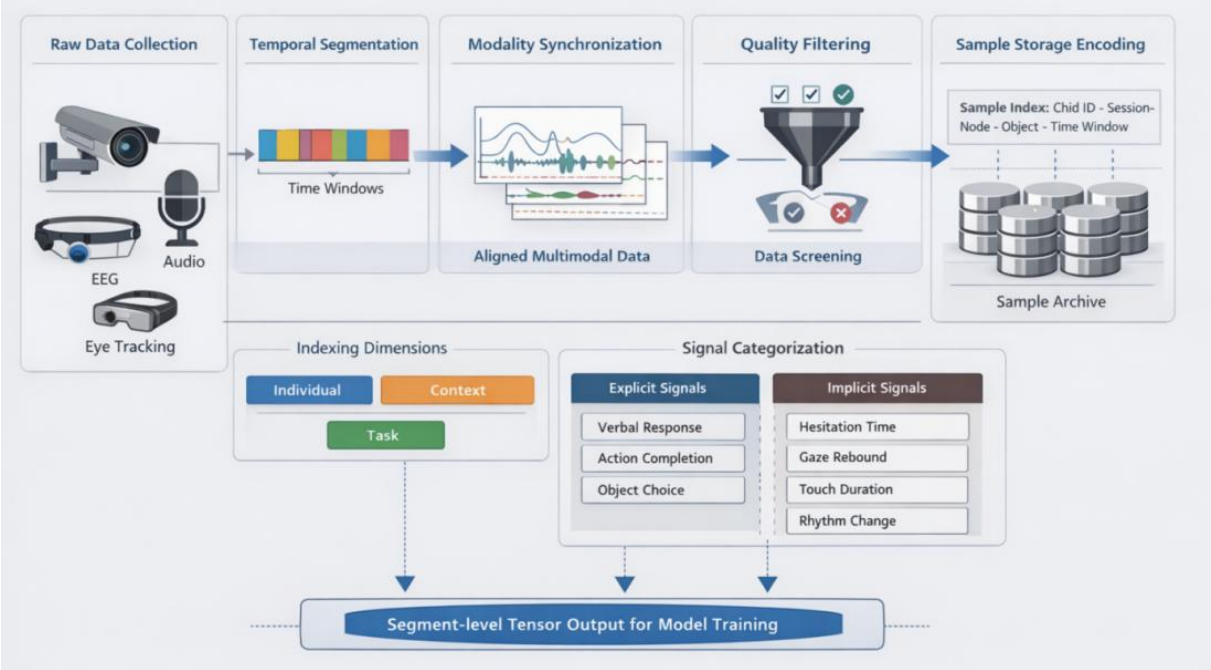

Figure 1: Structure diagram of multimodal data flow and acquisition mechanism in natural aesthetic education scene

In order to organize visual frames, speech segments, action sensing values, gaze regions and environmental object labels in the same activity window into a unified input, the original multimodal segments are defined as follows.

$$\mathcal{X}_i = \{(v_{i,t}, a_{i,t}, m_{i,t}, g_{i,t}, e_{i,t}, \tau_t) \mid t = 1, \ldots, T\} \tag{1}$$

Here, $\mathcal{X}_i$ represents the multimodal sequence of the i child in an activity segment, $v_{i,t}$, $a_{i,t}$, $m_{i,t}$, $g_{i,t}$, $e_{i,t}$ represent visual, speech, action, gaze, and environmental semantic features, respectively, and τ_t represents the timestamp. The function of this formula is to unify the heterogeneous inputs on the same time axis and provide a basic sample structure for subsequent cross-modal coding.

Considering that the sampling frequency and recording density of different acquisition terminals are not consistent, and the effective experience in natural activities depends on continuous time relationships, the system further maps non-equally spaced observations into a unified time window representation:

$$\tilde{x}_{i,t}^{r} = \frac{1}{|\Omega_t^r|} \sum_{\omega \in \Omega_t^r} x_{i,\omega}^{r}, \quad r \in \{v, a, m, g, e\} \tag{2}$$

Here, $\tilde{x}_{i,t}^{r}$ denote the resampling result of mode r on window t, Ω_t^r denotes the set of

observations falling into this window, and $|\Omega_t^r|$ denotes the set size. This formula is used to unify the high frequency and low frequency signals into a fixed window scale, and reduce the bias caused by the difference of sampling density.

Due to the quality fluctuations caused by occlusion, background noise, local out-of-focus and viewpoint shift, the system introduces quality gating to suppress unstable inputs before loading the samples into the library:

$$q_{i,t}^r = \sigma(w_q^{r\top} f_{i,t}^r + b_q^r), \quad \hat{x}_{i,t}^r = q_{i,t}^r \cdot \tilde{x}_{i,t}^r \tag{3}$$

Here, $q_{i,t}^r$ represent the quality score of modality r on time window t, and $f_{i,t}^r$ represent the quality discriminant statistical vector. This formula ensures that low definition and low discernible signals are suppressed before entering the training set.

After the above processing, the system establishes a stable multimodal data entry. The acquisition mechanism not only preserves the scene continuity of natural aesthetic education activities, but also ensures the calculation requirements of subsequent deep models for sample temporal consistency and modal correspondence. This section completes the data organization at the lowest level of the system, which not only serves the subsequent feature extraction, but also determines the real-time performance and deployment stability of the whole analysis system.

3.2 Emotion cognitive feature extraction and Cross-modal semantic coding

Emotional cognition in natural aesthetic education is not represented by a single emotional label, but by expression tension, voice fluctuation, dwell depth, gait rhythm and object look back. Based on this feature, the system does not directly use raw pixels, waveforms and sensing curves at this stage, but extracts local representations of vision, speech, action, gaze and environmental objects respectively, and then obtains a unified emotional cognitive vector through cross-modal semantic coding. The visual encoder is responsible for extracting eye-mouth co-variation, facial muscle strength, hand contact movements, and body orientation. The speech encoder is responsible for extracting the pitch contour, energy peak, pause structure and emotional word intensity. The action encoder is responsible for extracting speed changes, steering patterns, stopping rhythms, and approach behavior. The gaze encoder is responsible for extracting the hot spot center, stay span and backsweep frequency. The environment encoder encodes plants, water bodies, stone leaves, light and texture objects into semantic labels. After normalization and dimension compression, each modal representation enters a shared space, thus forming an emotional cognitive state with scene context.

Fig. 2 shows the complete path from unimodal encoding to shared semantic generation. The five types of inputs first complete local representation learning in their respective deep encoders, and then establish semantic complementarity through the cross-modal alignment module. When young children exhibit a raised tone, prolonged gaze, and slower gait when observing leaf textures, these signals will be viewed by the system as composite expressions of the same experience event, rather than as mutually independent segment inputs.

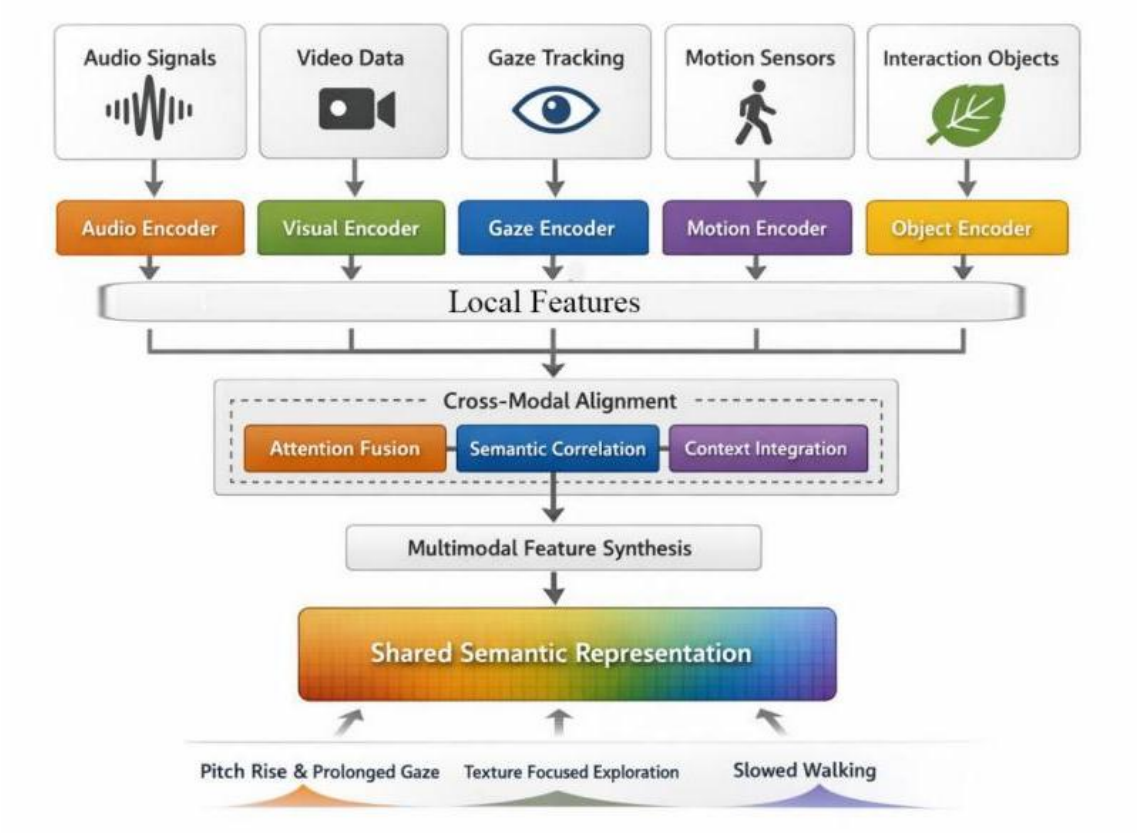

Figure 2: Flowchart of emotion cognitive feature extraction and cross-modal semantic encoding

In order to learn the local semantic features of each modality in natural activities separately and retain enough information density for subsequent sharing coding, this paper constructs an independent coding function for each type of modality:

$$z_{i,t}^{r} = \phi_r(s_{i,t}^{r}; \Theta_r), \quad r \in \{v, a, m, g, e\} \tag{4}$$

Here, $s_{i,t}^{r}$ represent the input of the r modality at time window t, $\phi_r(\cdot)$ represent the deep encoder of the corresponding modality, Θ_r represent the parameters of the encoder, $z_{i,t}^{r}$ represent the encoded local semantic vector. This formula is used to compress inputs from different sources and distributions into a comparable hidden space.

Considering that the intensity of emotional expression in natural scenes is not balanced, simply concatenating each modal vector is easy to dilute the key signal, so the system further introduces modal attention weights to form a joint representation:

$$\alpha_{i,t}^{r} = \frac{\exp(q_{i,t}^{\top} k_{i,t}^{r} / \sqrt{d})}{\sum_{u} \exp(q_{i,t}^{\top} k_{i,t}^{u} / \sqrt{d})}, \quad h_{i,t} = \sum_{r} \alpha_{i,t}^{r} z_{i,t}^{r} \tag{5}$$

Here, $\alpha_{i,t}^{r}$ represents the attention weight of the r modality over time window t, $q_{i,t}$ represents the query vector, $k_{i,t}^{r}$ represents the key vector of the r modality, and $h_{i,t}$ represents the joint semantic state. This formula can automatically improve the importance of visual and gaze signals when speech is weak but gaze is stable and movement is slow.

In order to ensure that the joint representation not only emphasizes the local strong signal, but also maintains the semantic consistent direction of the same experience event in different

modalities, the consistency constraint is added:

$$\mathcal{L}_{\text{sem}} = \sum_{t=1}^{T} \sum_{r<u} \left(1 - \frac{z_{i,t}^{r\top} z_{i,t}^{u}}{\|z_{i,t}^{r}\|_2 \|z_{i,t}^{u}\|_2} \right) \tag{6}$$

Here, $\mathcal{L}_{\text{sem}}$ denotes the semantic consistency loss, $z_{i,t}^{r\top} z_{i,t}^{u}$ denotes the inner product of different modal vectors, and $\|\cdot\|_2$ denotes the two-norm. This formula makes the description of the same experience event by vision, speech and movement keep the same direction.

Since the natural object itself will have a modulation effect on children's emotional cognitive strength, the system further introduces environmental semantic gating, so that the object attributes directly participate in the joint representation calculation:

$$c_{i,t} = \sigma(W_c e_{i,t} + b_c) \odot h_{i,t} \tag{7}$$

where $c_{i,t}$ represents the emotional cognitive representation after introducing object semantic modulation, W_c and b_c are parameters, $\sigma(\cdot)$ is the gating function, and $e_{i,t}$ represents the environment semantic vector. This formula enables the system to consider both the subjective response of the young child and the semantic properties of the current natural object during the encoding process.

Through the above steps, the system generates high quality shared semantic vectors at this stage that can be directly invoked by subsequent temporal modeling, behavior inference, and online feedback. What is completed here is not ordinary feature compression, but the computational representation of emotional cognition constructed for natural aesthetic education scene, which provides the core input for the deep learning backbone of the whole system.

3.3 Deep Learning Driven System architecture

After multi-modal acquisition and semantic encoding, the system needs to transform the segment-level representation into a unified deep model that can support multi-task prediction at the same time. To this end, this paper adopts the structure of "shared backbone network + multi-task output head". The shared backbone consisted of a time and position encoding layer, a gated timing layer, a global attention aggregation layer and a state cache layer, which were used to model the continuous state changes during children's activities. The multi-task output heads correspond to experience participation prediction, emotional state recognition and scene semantic matching scoring respectively. The state cache layer sends the result of the current segment to the online feedback module. The advantage of this structure is that multiple tasks can share features in the same representation space, which not only reduces parameter redundancy, but also utilizes the correlation between tasks to improve the overall stability. In the training phase, the joint loss is used to optimize all modules, and in the deployment phase, the distilled lightweight weight is used to realize the edge-end inference, so as to balance the accuracy and delay.

Fig. 3 shows the complete flow from shared representation input to multitask output generation. The emotional cognitive vector output by the coding layer is firstly compensated by time and position, and then the history dependence is extracted by the gated temporal unit, and then the activity-level context representation is formed in the global attention layer. Finally, the output is sent to three task heads at the same time. The key to this figure is not the number of stacks, but the fact that the backbone representation can serve classification, regression, and matching tasks simultaneously, allowing the system to obtain a complete state determination in a single forward pass.

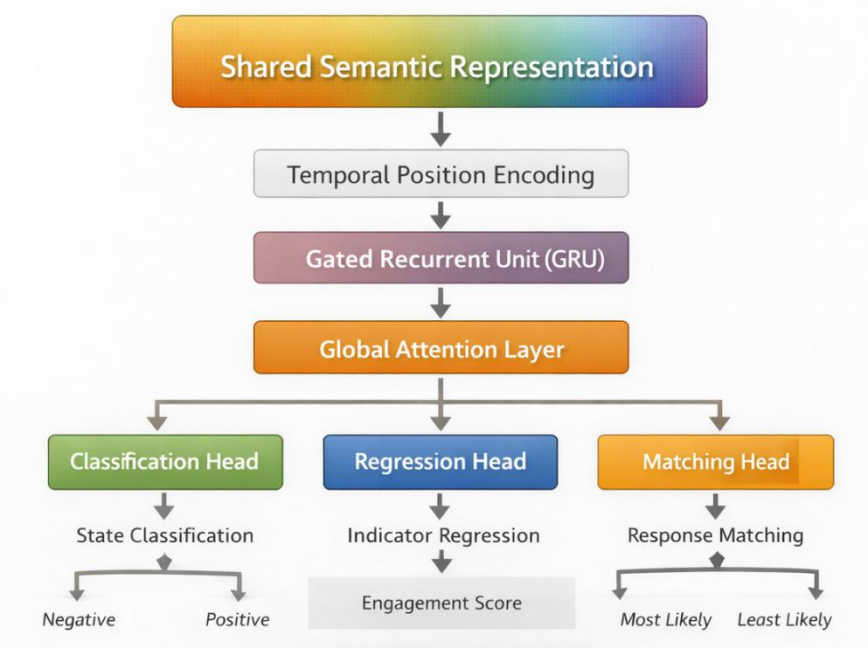

Figure 3: Deep learning driven system architecture diagram

In order to allow the model to explicitly distinguish the time location of the different phases of activity initiation, continuous observation, repeated contact, and end fall, the system first adds a location encoding before the input trunk:

$$p_{t,2k} = \sin\left(\frac{t}{10000^{2k/d_p}}\right), \quad p_{t,2k+1} = \cos\left(\frac{t}{10000^{2k/d_p}}\right) \tag{8}$$

Here, $p_{t,\cdot}$ denotes the temporal position vector, k denotes the dimension index, and d_p denotes the length of the position vector. This formula enables the system to recognize the stage position under similar expression or action pattern.

In order to model continuous state propagation inside segments and suppress transient noise while preserving important historical information, this paper uses gated recursion to update hidden states:

$$s_t = u_t \odot s_{t-1} + (1 - u_t) \odot \tanh(W_s c_t + U_s(r_t \odot s_{t-1}) + b_s) \tag{9}$$

Here, s_t represents the hidden state at time t, u_t and r_t represent the update gate and reset gate, respectively, and W_s, U_s and b_s are parameters. This equation is used to retain stable experience cues and suppress local disturbances during activity.

In order to extract the key moments that best represent the quality of experience from the entire activity, rather than averaging over all Windows, temporal attention aggregation is used to build a global representation:

$$\beta_t = \frac{\exp(w_\beta^\top \tanh(W_\beta s_t + b_\beta))}{\sum_{j=1}^{T} \exp(w_\beta^\top \tanh(W_\beta s_j + b_\beta))}, \quad c = \sum_{t=1}^{T} \beta_t s_t \tag{10}$$

Here, β_t denotes the time-step weight, c denotes the activity-level context vector, and W_β, w_β, and b_β are parameters. This formula can automatically highlight high-value moments such as first contact, vocalization after sustained gaze, or repeated look back at the

object.

In order to simultaneously output experience participation probability, emotional state distribution and scene semantic matching score on the same backbone, the system maps the global context to different task Spaces:

$$\hat{m} = \sigma(w_m^\top c + b_m), \quad \hat{y} = \mathrm{Softmax}(W_y c + b_y), \quad \hat{q} = w_q^\top c + b_q \tag{11}$$

Here, $\hat{m}$ represents the experience participation probability, $\hat{y}$ represents the emotional state distribution, $\hat{q}$ represents the scene semantic matching score, and w_m, W_y, w_q and the corresponding bias are the task head parameters. This formula ensures that three types of output can be completed in one forward propagation.

In order to prevent a certain task from generating too strong traction on the backbone representation, the system further defines a joint optimization objective:

$$\mathcal{L}_{\mathrm{all}} = \lambda_1 \mathcal{L}_{\mathrm{eng}} + \lambda_2 \mathcal{L}_{\mathrm{emo}} + \lambda_3 \mathcal{L}_{\mathrm{match}} + \lambda_4 \|\Theta\|_2^2 \tag{12}$$

Here, $\mathcal{L}_{\mathrm{eng}}$ represents the participation prediction loss, $\mathcal{L}_{\mathrm{emo}}$ represents the sentiment classification loss, $\mathcal{L}_{\mathrm{match}}$ represents the matching regression loss, λ_1 to λ_4 are the balance coefficients, Θ represents the set of parameters. With this objective function, the system simultaneously takes into account recognition accuracy, matching quality and parameter stability during training.

Through this architecture design, the system realizes a complete calculation path from shared representation to multi-task inference in the same deep model, which not only meets the requirements of technical issues for model unity and verifiability, but also provides a reusable state representation for the next part of behavior influence relationship modeling.

3.4 Behavior influence relationship modeling and inference strategy

In natural aesthetic education activities, children's behavior is not an isolated action, but the result of object stimulation, spatial path and emotional cognition. Judging "participation or not" simply based on the action sequence is difficult to explain why the behavior occurs, and it is difficult to reveal which kinds of objects are more likely to cause stable exploration or short-term avoidance. Based on this understanding, this paper abstracts the activity session as a heterogeneous relationship graph with time attribute. The nodes include children nodes, natural object nodes, scene area nodes and activity event nodes, and the edges represent the relationships such as contact, stay, look back, move, sound and interaction switch. The system uses the graph structure to uniformly express the association link of "object attribute-cognitive state-behavior result", and then calculates the behavior influence strength by graph aggregation and state score, and finally outputs the behavior category and trajectory evaluation result.

Fig. 4 shows the complete process from node construction, relationship edges to behavior influence inference. In each time slice, child nodes form dynamic connections with object nodes, region nodes and event nodes. The system first calculates edge weights, then performs neighborhood aggregation, and then outputs object influence strength, trajectory consistency and behavior category. The key point is that actions are not directly classified as end labels, but instead are calculated in an interpretable network of relationships, which allows the system to account for the actual contribution of a certain class of objects, a certain path, or an active node to a change in behavior.

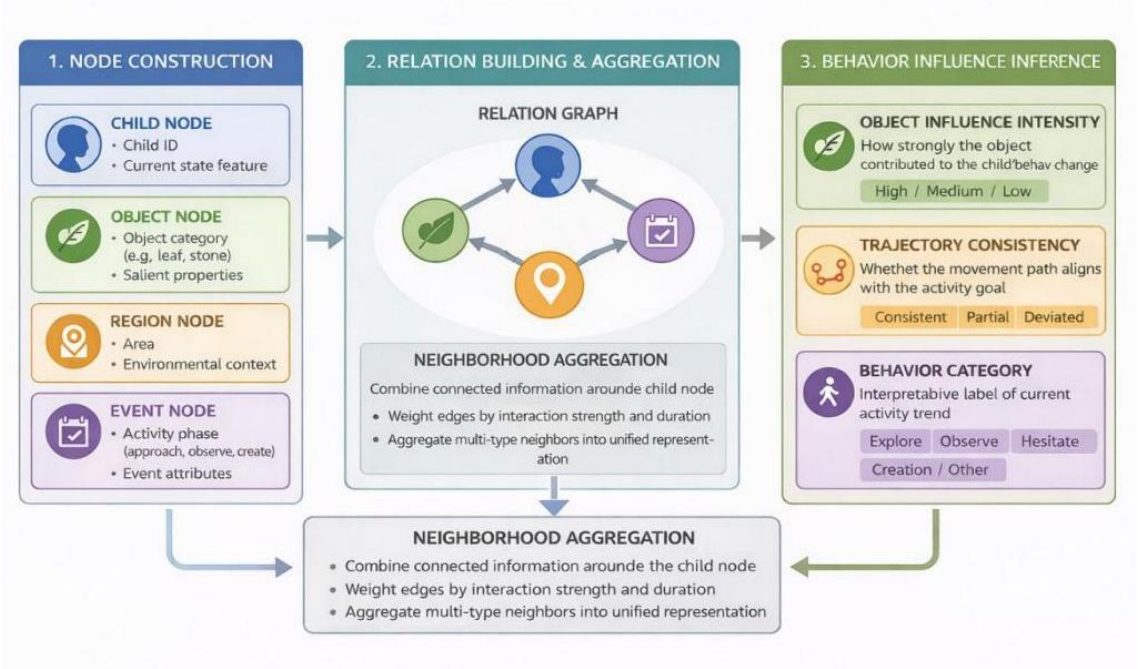

Figure 4: Behavioral influence relationship modeling and inference strategy diagram

In order to describe the contribution strength of different objects, regions and events to the current behavior state, and to enable the model to distinguish the actual role of different relationships on heterogeneous graphs, the system first defines the dynamic edge weights between nodes:

$$a_{uv}^{(t)} = \frac{\exp(r_{uv}^{\top} M[n_u^{(t)} \| n_v^{(t)} \| c_{uv}])}{\sum_{k \in \mathcal{N}(u)} \exp(r_{uk}^{\top} M[n_u^{(t)} \| n_k^{(t)} \| c_{uk}])} \tag{13}$$

Here, $a_{uv}^{(t)}$ represents the relationship weight from node u to node v at time t, $n_u^{(t)}$ and $n_v^{(t)}$ represent node states, c_{uv} represents edge attributes, and $r_{uv}$and M are parameters. This equation is able to distinguish between semantically different relationship strengths such as "short stay" and "continuous contact".

After obtaining the dynamic edge weights, the system compressions the influence of objects, regions, and active events in the neighborhood into a contextual representation of the current node to construct the graph state needed for behavior inference:Here, $g_u^{(t)}$ represents the aggregate representation of node u at time t, $\rho(\cdot)$ is the nonlinear activation function, and W_g and B_g are parameters. This formula enables the child node to absorb the contextual influence of surrounding objects and events when updating its own state.

In order to measure the persistence and phase consistency of a behavior trajectory during an activity, rather than just whether there is an action at a certain instant, the system further defines a trajectory consistency score:

$$\Gamma_i = \frac{1}{L_i} \sum_{l=1}^{L_i} [\eta_1 \cos(g_{i,l}, g_{i,l-1}) + \eta_2 \Delta p_{i,l} + \eta_3 \Delta d_{i,l}] \tag{15}$$

Here, Γ_i represents the consistency score of the i trajectory, L_i represents the trajectory length, $\Delta p_{i,l}$ represents the position change, $\Delta d_{i,l}$ represents the stay depth change, and η_1

to η_3 are the weights. This formula is used to determine whether children are in stable exploration, continuous observation, or frequent switching and avoidance states.

Through this modeling, the system is able to recognize not only what the child does, but also how this behavior is shaped by both the environment and the emotional state. After this part is completed, the system has the conditions to send the behavior results back to the activity site for adaptive adjustment.

3.5 Closed loop of online feedback and Experience adaptation

The online feedback module is responsible for the closed-loop function of "reading the state in real time, generating the adjustment action, observing the action effect and updating the subsequent strategy". The system receives experience participation probability, emotional state distribution, scene semantic matching score and behavior influence category after each activity window, and then the feedback decision maker determines whether the current activity is maintained, fine-tuned or switched. Fine-tuning actions include extending object contact time, adjusting the sequence of material presentation, enhancing voice cues, adding touchable objects, switching viewing positions, or changing the guiding rhythm. In order to prevent excessive adjustment caused by a single abnormal window, the system smoothen and caches the multi-task output before entering the policy calculation, and then generates the action probability by combining the change trend of recent Windows. At the same time, the system retains the effect label after each adjustment for subsequent incremental learning and parameter correction. In this way, the system is not just an offline recognizer, but an online intelligent terminal that can continuously update judgments and suggestions with the activity process.

Fig. 5 illustrates the complete closed loop of "model output -state smoothing -policy generation -activity adjustment -result writeback -parameter update". The decision maker decides whether to keep the original path, fine-tune the activity sequence, or switch to a new object or region based on the current state. The goal of fine-tuning is not to simply improve the number of participation, but to form a more stable positive consistency of emotional cognition, semantic matching and behavior trajectory. The key point in the figure is that all feedback actions are timestamped and effect labeled, which enables the system to gradually accumulate learnable policy experience as it runs.

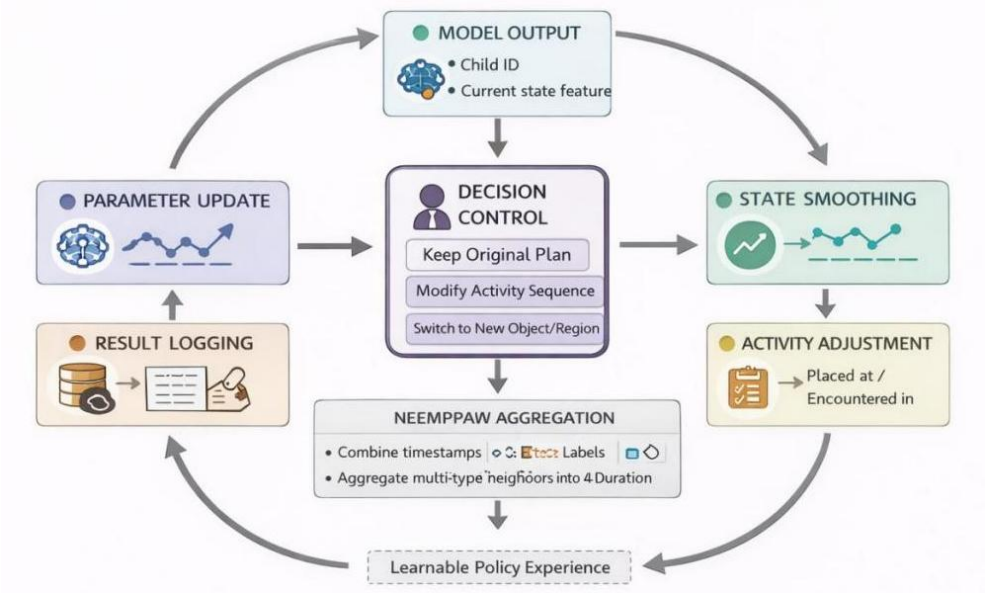

Figure 5: Closed-loop graph of online feedback and experience adaptation

In order to organize participation probability, sentiment distribution, semantic matching score and behavior influence category into a unified state input that can be directly used by the feedback decision maker, the system first defines the online state vector:

$$o_t = [\hat{m}_t, \hat{y}_t, \hat{q}_t, \hat{b}_t, \Delta\hat{m}_t, \Delta\hat{q}_t] \tag{16}$$

where o_t represents the online state vector at time t, $\hat{m}_t$ represents the experience participation probability, $\hat{y}_t$ represents the emotion distribution, $\hat{q}_t$ represents the scene semantic matching score, and $\hat{b}_t$ represents the behavior influence category. This formula integrates multiple task outputs into a single feedback space.

Since there are many short-term disturbances in natural scenes, the system needs to suppress occasional anomalies while preserving the change trend, so exponential smoothing is used to generate the pre-feedback state:

$$\tilde{o}_t = \lambda o_t + (1-\lambda)\tilde{o}_{t-1} \tag{17}$$

Here, $\tilde{o}_t$ represents the smoothed state vector and λ represents the current window weight. This equation reduces the pull of a single abnormal window on strategy switching and makes the online regulation more in line with the rhythm of continuous activities.

In order to select the feedback mode that best matches the current state among various actions such as maintaining, enhancing guidance, adjusting order, and switching nodes, the system defines the policy network output:

$$\pi(a_t \mid \tilde{o}_t) = \frac{\exp(w_{a_t}^{\top}\tilde{o}_t)}{\sum_{a\in\mathcal{A}}\exp(w_a^{\top}\tilde{o}_t)} \tag{18}$$

Here, a_t represents the current feedback action, $\mathcal{A}$ represents the action set, and w_a represents the parameters of each action. This formula makes the feedback no longer depend on the fixed threshold, but adaptively generate the action probability according to the current state distribution.

In order to evaluate whether an adjustment actually leads to a better experience state and convert the adjustment effect into a learnable signal, the system further defines an immediate reward function:

$$R_t = \xi_1(\hat{m}_{t+1} - \hat{m}_t) + \xi_2(\hat{q}_{t+1} - \hat{q}_t) + \xi_3\Delta\Gamma_{t+1} \tag{19}$$

Here, R_t represents the immediate reward brought by the action a_t, $\hat{m}_{t+1} - \hat{m}_t$ represents the engagement change, $\hat{q}_{t+1} - \hat{q}_t$ represents the semantic matching change, $\Delta\Gamma_{t+1}$ represents the trajectory consistency improvement, and ξ_1 to ξ_3 are the weights. This formula can quantify the actual effect of feedback action as the basis for subsequent update.

Through this closed-loop mechanism, the system can continuously adjust the strategy according to the current activity state, and write the adjusted results back to the subsequent training and inference link, thus forming a true online adaptive analysis system.

4 Results and discussion

4.1 Experience participation prediction accuracy analysis

Experience engagement prediction is used to test the system's ability to discriminate children's engagement states in nature observation, object contact, path following and immediate

interaction. In the experiment, the activity segments were divided into three categories: high participation, medium participation and low participation, and the manual review label was used as a benchmark for control analysis. As shown in Fig. 6, the recognition boundaries of the proposed system on the two categories of high participation and medium participation are relatively clear, and the high participation samples are mainly concentrated in the diagonal region, indicating that the model has strong discrimination ability for behaviors such as continuous gaze, repeated contact and active response. The misclassification of low-participation samples mainly occurs between short-term hesitation and mild dissociations, which is related to the segment characteristics of small movement amplitude and weak speech output in natural scenes, but not yet completely out of activity.

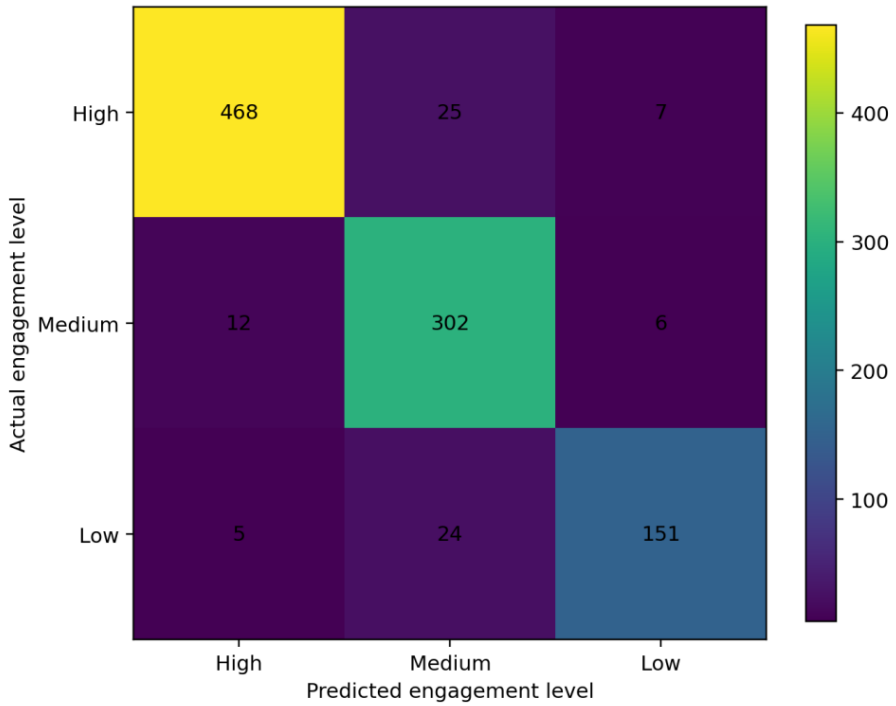

Figure 6: Heat map of confusion matrix for experience engagement prediction

Compared with Baseline-CNN, the proposed system significantly reduces the category jump caused by single frame fluctuation after continuous window aggregation. Compared with the AV-Transformer, the proposed system is more stable in the recognition of the forward and backward correlation features such as "stay before contact - look back after contact" after introducing the environmental semantics and gaze trajectory. The test results show that the accuracy of experience participation prediction of our system reaches 92.1%, Macro-F1 reaches 91.4%, and the recall rate of high participation category reaches 93.6%. This result shows that the multimodal joint modeling can completely retain the process of children's investment in natural aesthetic education activities, rather than only capturing the final action results.

It can be further seen from the misclassified positions in the heatmap that the errors are mainly concentrated in two time Windows before and after the transition of the active node. At this time, children often appear at the same time signals such as short stop, gaze deviation and speech attenuation. If we only rely on a single mode in vision or action, it is easy to misjudge samples that are about to enter a high participation state as moderate participation. The proposed system weakens this effect through state caching and environment semantic compensation, so it still maintains high stability in the node switching phase. On the whole, the experience participation prediction module has a good forward-looking recognition ability, which can

provide a reliable basis for subsequent individualized adjustment.

4.2 Analysis of emotional state recognition accuracy

Affective state recognition is used to measure the ability of the system to distinguish four types of states: pleasure, calm, hesitation and avoidance. This part of the experiment focuses on the collaborative contributions of expression, speech, gaze and action signals in different activity segments. As shown in Fig. 7, the recognition accuracy of the proposed system is high on both pleasure and calm classes, and the overall boundary is significantly better than that of the control model although there is still a small amount of overlap between hesitation and avoidance. The reason is that cross-modal semantic encoding not only preserves facial and acoustic cues, but also incorporates object attributes, scene locations and gaze pauses into the representation space, which enables the system to distinguish between "quiet observation" and "emotional avoidance", two states with similar surface actions but different semantics.

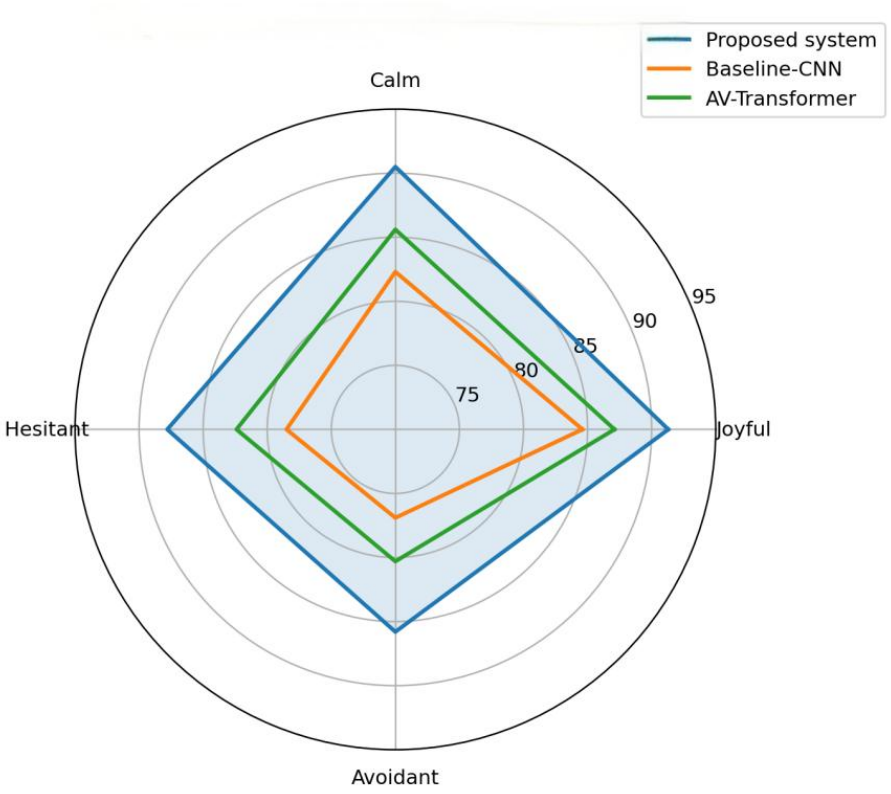

Figure 7: Radar chart of affective state recognition

The experimental results show that the overall recognition accuracy of our system is 89.7%, and the weighted F1 rate is 88.9%. The accuracy of happy state is 91.3%, and the recall rate of calm state is 90.5%. Compared with Baseline-CNN, the cross misclassification of the proposed system in the hesitation and avoidance categories is significantly reduced. Compared with the AV-Transformer, the recognition of the proposed system on calm states is more stable, indicating that the environmental semantic vector has an obvious compensation effect on the determination of low-intensity emotional states. The distribution profile in the radar chart also shows that the performance of the proposed system on the four categories of states is more balanced, and there is no excessive bias towards the high-frequency category.

In terms of category features, hesitancy is usually accompanied by short observations, low amplitude movements and intermittent gaze pauses, while avoidance is more likely to be

manifested by gaze disengagement, path deflection and speech silence. If we judge only by expression intensity, these two types of states are easily compressed into the same range. Our system preserves these subtle differences by jointly encoding expression, gaze, and action, so the emotion recognition results are closer to the real state evolution in natural activities. This indicates that the emotional state module already has the ability to provide fine-grained input for scene matching and online feedback.

4.3 Scene semantic matching score analysis

The scene semantic matching score is used to evaluate the degree of agreement between the object proposal, the active node recommendation output by the system and the current state of the child. In the experiment, the system sends the semantic vector of natural objects and the current emotional cognitive representation of children into the matching head, and calculates the fit between the recommended content and the current experience state. As shown in Fig. 8, the high matching samples generated by the proposed system are clustered more closely in the shared space, while the low matching samples are more distributed in the area where the object attribute deviates greatly from the current state. This indicates that the system can not only identify the emotional and behavioral state of the toddler at the moment, but also find a natural object or activity path that is more suitable for the current experience rhythm.

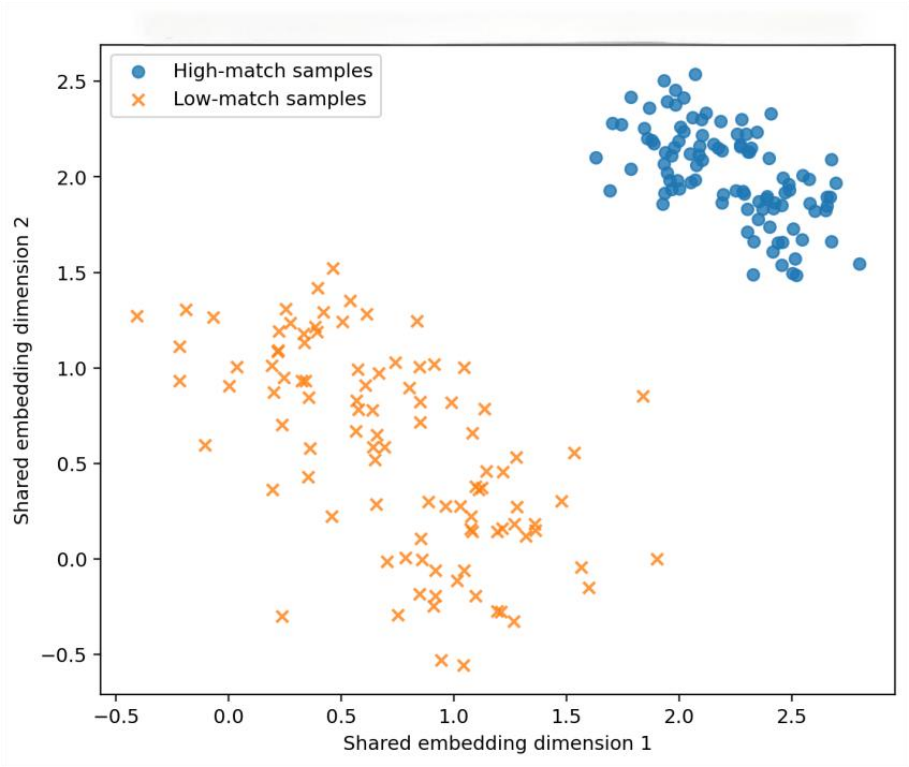

Figure 8: Scatter plot of 2D embeddings for scene semantic matching

Compared with the baseline model that only uses visual semantics, the proposed system can better judge whether children truly maintain aesthetic interest in an object after adding speech and gaze. Compared with the control method that only uses behavior sequence, our system avoids misclassifying mechanical touch or passive dwell as high-quality experience response. The average semantic matching score of the final experiment is 87.9%, and the Top-2 matching hit rate is 93.1%. From the scatter plot distribution, it can be seen that the high-matching segments usually have the three characteristics of stable fixation, repeated object contact and

positive acoustic feedback at the same time, while the low-matching segments mostly appear in the scenes where the object switches too fast or the environmental stimulus is inconsistent with the current cognitive rhythm.

Further analysis of the boundary shape of the embedding space shows that our system retains the continuous correspondence between object features and individual states in the shared space, so the recommendation results not only have semantic consistency, but also have good experience coherence. Compared with the static object ranking method, the proposed method is more suitable for the continuously changing open environment of natural aesthetic education activities. The results show that the scene semantic matching module can not only be used to evaluate the output quality of the system, but also directly serve the subsequent recommendation ranking and adjustment decisions.

4.4 Online inference delay analysis

The online inference delay reflects the time it takes for the system from receiving the multimodal window to output the participation probability, sentiment category, semantic match, and behavioral influence labels. Table 2 shows that the average delay of the proposed system is 84 ms, which is significantly lower than the two groups of models with higher parameter numbers, and also lower than GraphFusion without light distillation. This result is consistent with the shared backbone structure and the edge inference design in Section III. For natural aesthetic education activities, if the response delay is too long, even if the semantic of the system suggestion is correct, it is difficult to form a synchronous relationship with the current experience state, thus weakening the practical role of the feedback.

Table 2: Online inference delay analysis

Method	Average Latency / ms	Maximum Latency / ms	Minimum Latency / ms	Latency Standard Deviation
Proposed System	84	109	71	8.6
Baseline-CNN	126	158	103	12.4
AV-Transformer	143	181	117	15.1
GraphFusion	97	128	82	9.7

From the perspective of specific indicators, the maximum delay of the system in this paper is controlled within 109 ms, and the standard deviation of the delay is 8.6, indicating that the model not only has a fast average speed, but also has a small reasoning fluctuation. Compared with Baseline-CNN, although the multimodal input and scene semantic modeling are added to the proposed system, the overall response efficiency is better due to the unified backbone structure and shared task head. Compared with AV-Transformer, the proposed system significantly reduces the inference time while maintaining high accuracy, which indicates a good balance between joint encoding and lightweight deployment.

Further observation of the reasoning records under different active nodes shows that the delay fluctuations mainly appear in the scene switching and object dense stages, but the overall variation range is small, and there is no long time blocking phenomenon. For field deployment, this feature is more valuable than the single minimum delay, because the stable response rhythm can ensure that the system prompt is in sync with the current behavior of the child, and avoid the misalignment between the advice output and the real experience stage. In summary, the system in this paper has met the requirements of near-real-time analysis in natural scenes, and has high engineering usability.

4.5 Effectiveness analysis of individualized adjustment

The effectiveness of individualized adjustment is used to measure whether the actions such as object switching, sequence adjustment, duration extension and cue enhancement given by the system according to the current state can guide children from a low-quality experience state to a more stable positive state. The experiment divided the window before and after adjustment into four categories: avoidance, hesitation, calm observation and active exploration, and analyzed the state distribution before adjustment, state transition matrix and state distribution after adjustment. As shown in Fig. 9, after adjustment, the proposed system makes 31.4% of hesitant samples transfer to calm observation, and makes 24.7% of calm observation samples further transfer to active exploration, and the overall migration direction is better than the fixed rule strategy.

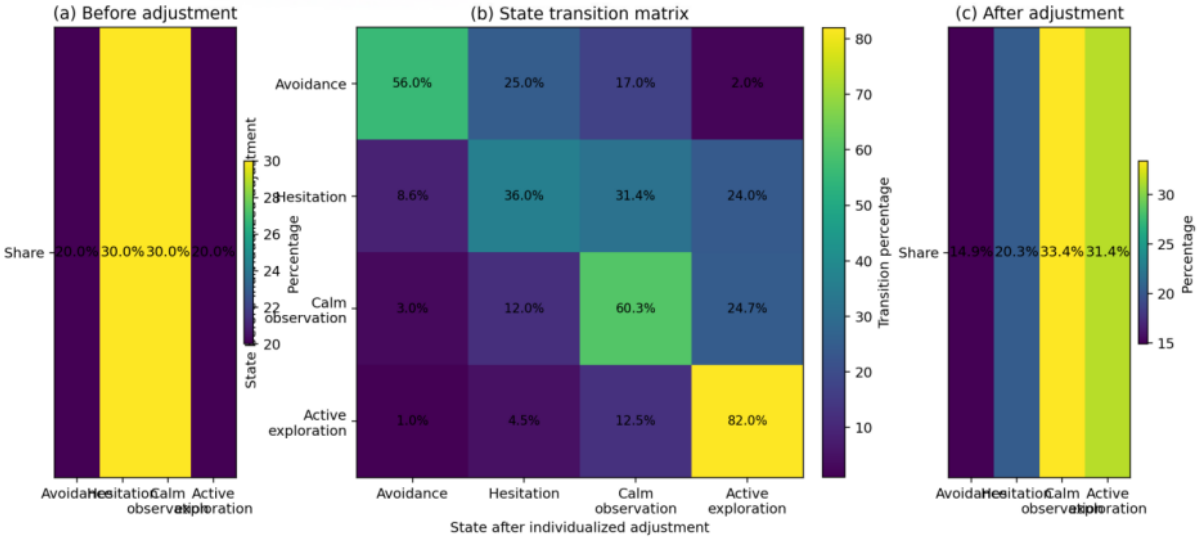

Figure 9: Triplet heatmaps of state distribution and transition before and after individualized adjustment

This result shows that the feedback decision-maker does not simply improve the interaction density, but makes targeted fine-tuning according to object attributes, emotional states and trajectory consistency. Compared with the rule-driven approach, the proposed system can grasp when the object contact needs to be prolonged, when the presentation order needs to be adjusted, and when the scene node should be switched to a more suitable scene node for the current state. The experimental results show that the effectiveness of individualized adjustment of the proposed system reaches 86.2%, which is 11.8 percentage points higher than that of the fixed rule method.

It can be seen from Figs. 9(a) to 9(c) that the improvement of the proposed system is most obvious in the two links of "hesitation-calm observation" and "calm observation - active exploration", while the prudent adjustment mode is maintained in the cross-level transfer of "avoidance - active exploration". This transfer structure is more in line with the natural activity rhythm, which also indicates that the feedback logic of the model is phased and continuous. It can be seen that the individualized adjustment module has execution value and can provide support for experience response gain.

4.6 Response gain analysis of natural aesthetic education experience

The response gain of natural aesthetic education experience reflected the improvement of children's more stable fixation, longer object stay, more positive voice feedback and higher repeated observation intention in continuous activities after the intervention of the system. Table 3 shows that the system in this paper is at the highest level in the three indicators of response gain, retention rate and average experience score, in which the response gain reaches 18.6% and the retention rate reaches 88.4%, indicating that the object suggestions and path adjustments generated by the system can continuously enhance children's aesthetic input, rather than only produce local fluctuations in a short window.

Table 3: Response gain analysis of natural aesthetic education experience

Method	Response Gain / %	Retention Rate / %	Average Experience Score	Positive Feedback Growth / %
Proposed System	18.6	88.4	91.7	21.3
Baseline-CNN	9.4	74.8	79.1	10.7
AV-Transformer	12.1	78.3	82.5	14.2
GraphFusion	15.3	83.6	87.0	17.4

Compared with Baseline-CNN, the proposed system has a larger improvement in positive feedback growth, which indicates that the multimodal semantic encoding can capture the current state more accurately and output more suitable regulation suggestions. Compared with AV-Transformer, the advantage of the proposed system in retention rate is more obvious, indicating that the introduction of object attributes and environmental semantics helps to maintain the continuous engagement of young children in subsequent segments. Although GraphFusion performs well on multi-source information integration, its response gain and positive feedback growth are still lower than the proposed system, indicating that its utilization of real-time state changes is not sufficient.

According to the relationship between each index, response gain and retention rate showed a positive correlation trend, indicating that the system output not only enhanced the immediate response intensity, but also improved the possibility of children to maintain aesthetic investment in subsequent activities. Compared with the recommendation strategy that relies on static ranking, the proposed system can better reflect the continuous coupling relationship between "current state-object feature-subsequent feedback". On the whole, the response gain analysis of natural aesthetic education experience shows that the proposed method is not only suitable for state recognition, but also has the ability to continuously enhance the quality of experience.

4.7 Analysis of behavior influence trajectory recognition

The behavior influence trajectory recognition is used to test whether the system can accurately distinguish the four types of trajectory structures of "active exploration, stable observation, short-term hesitation, and avoidance exit" in continuous activities, and to determine the specific effects of object stimuli and path transitions on behavior changes. As shown in Fig. 10, the density distribution of the proposed system is more concentrated on the two categories of active exploration and stable observation, and the boundaries are also clearer on the two categories of short-term hesitation and avoidance exit. Compared with the method using only action sequences, the proposed system incorporates object semantics, emotional state and path changes into the analysis, so that it can more accurately distinguish the two seemingly similar but semantically different behavior trajectories of "stable stay after low-speed approach" and "avoidance exit after low-speed approach".

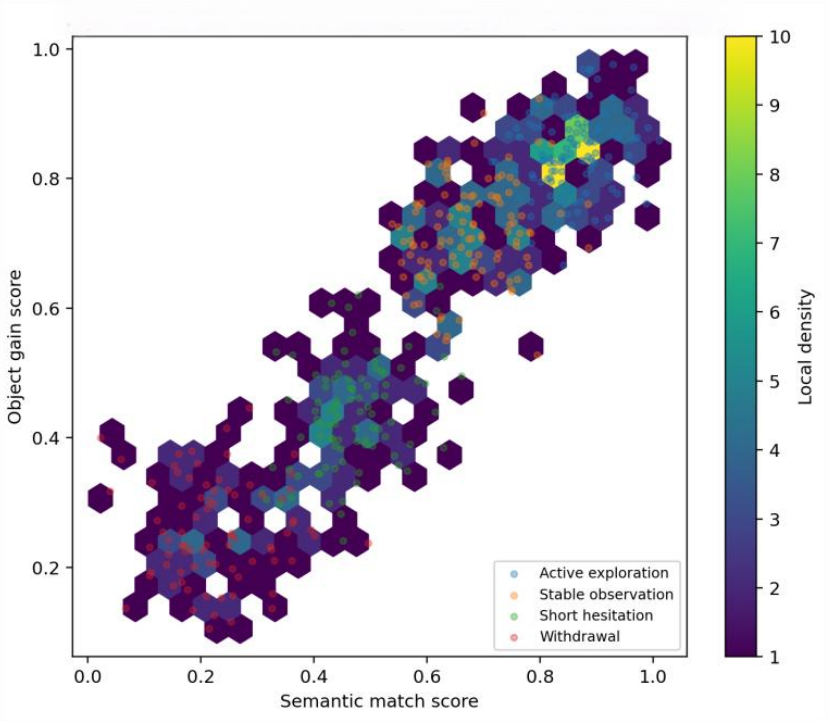

Figure 10: Density distribution map of behavioral influence trajectories

According to the local distribution, the active exploration trajectories were mainly concentrated in the high semantic matching and high object gain areas, indicating that when the color, texture and accessibility of natural objects were high consistent with children's current cognitive state, it was easier to form sustained observation, contact and repeated look back behaviors. The stable observation trajectory is more distributed in the middle and high matching region, which is manifested as smaller movement amplitude but longer fixation duration and dwell time. Short-term hesitation trajectories mainly appear in the position of object switching and scene node transition, and such segments are usually accompanied by brief stop, gaze deviation and movement rhythm weakening. The avoidance exit trajectory, on the other hand, was concentrated in areas of low matching and low object gain, indicating that young children were more likely to deviate from the activity path when the external stimulus was inconsistent with the current experience rhythm.

The experimental results show that the F1 of behavior influence trajectory recognition of our system reaches 90.2%, which is 10.4 percentage points higher than that of GraphFusion, indicating that graph structural relationship modeling indeed enhances the model's ability to understand the process of behavior formation, rather than only improving the end classification results. Further combining the trajectory density distribution, it can be seen that the distinction of the trajectory boundary of the proposed system does not depend on the single action intensity, but is based on the joint expression of object attributes, emotional states and temporal behaviors, so it has better stability and interpretability in continuous scenes.

5 Conclusions and future work

The experimental results show that the constructed analysis system of children's natural aesthetic education experience maintains stable performance on a number of key indicators. The accuracy of experience participation prediction reaches 92.1%, the accuracy of emotional state recognition reaches 89.7%, the score of scene semantic matching reaches 87.9%, and the

average online reasoning delay is controlled at 84 ms. The effectiveness of individualized adjustment reached 86.2%, the response gain of natural aesthetic education experience reached 18.6%, and the F1 of behavior influence trajectory recognition reached 90.2%. The above results show that the joint modeling of vision, speech, action, gaze and environmental semantics can completely describe children's aesthetic engagement, emotional changes and behavioral responses in natural activities. In addition, the performance fluctuations of the current model in low speech output samples, short-time window switching samples and continuous operation scenarios of edge devices still need to be further compressed, and the feedback strategy parameters, state smoothing coefficients and action decision boundaries still need to be calibrated with more real activity data. At the same time, the current system is still mainly built on the limited activity scene and the established label system. The transfer ability of cross-season, cross-region and cross-material environments needs to be further verified, and the modal stability under complex outdoor interference conditions still needs to be refined. On the basis of maintaining the existing framework, the subsequent research can continue to introduce more fine-grained object attribute coding, lightweight edge deployment, cross-scene incremental training and interpretable calculation mechanisms, and promote the continuous application of the system in the digital analysis and intelligent feedback of natural aesthetic education under the premise of strictly following the principles of anonymization, authorization use and minimal collection.

Funding

Project Source: Hainan Provincial Higher Education Teaching Reform Research Project

Project Title: Research on the Innovative Paths of Integrating Ideological and Political Elements into the Skill - oriented Courses for Preschool Education Major

Project Number: Hnjg2025zc-110

References

[1] Middya A I, Nag B, Roy S. Deep learning based multimodal emotion recognition using model-level fusion of audio–visual modalities[J]. Knowledge-based systems, 2022, 244: 108580.

[2] Kumar A, Sharma K, Sharma A. MEmoR: A multimodal emotion recognition using affective biomarkers for smart prediction of emotional health for people analytics in smart industries[J]. Image and Vision Computing, 2022, 123: 104483.

[3] Mehta N K, Prasad S S, Saurav S, et al. Three-dimensional DenseNet self-attention neural network for automatic detection of student's engagement[J]. Applied Intelligence, 2022, 52(12): 13803-13823.

[4] Rathod M, Dalvi C, Kaur K, et al. Kids' emotion recognition using various deep-learning models with explainable ai[J]. Sensors, 2022, 22(20): 8066.

[5] Selim T, Elkabani I, Abdou M A. Students engagement level detection in online e-learning using hybrid efficientnetb7 together with tcn, lstm, and bi-lstm[J]. Ieee Access, 2022, 10: 99573-99583.

[6] Tang G, Xie Y, Li K, et al. Multimodal emotion recognition from facial expression and

speech based on feature fusion[J]. Multimedia Tools and Applications, 2023, 82(11): 16359-16373.

[7] Le H D, Lee G S, Kim S H, et al. Multi-label multimodal emotion recognition with transformer-based fusion and emotion-level representation learning[J]. Ieee Access, 2023, 11: 14742-14751.

[8] Lian H, Lu C, Li S, et al. A survey of deep learning-based multimodal emotion recognition: Speech, text, and face[J]. Entropy, 2023, 25(10): 1440.

[9] Pan B, Hirota K, Jia Z, et al. A review of multimodal emotion recognition from datasets, preprocessing, features, and fusion methods[J]. Neurocomputing, 2023, 561: 126866.

[10] Mocanu B, Tapu R, Zaharia T. Multimodal emotion recognition using cross modal audio-video fusion with attention and deep metric learning[J]. Image and vision computing, 2023, 133: 104676.

[11] Gupta S, Kumar P, Tekchandani R. A multimodal facial cues based engagement detection system in e-learning context using deep learning approach[J]. Multimedia Tools and Applications, 2023, 82(18): 28589-28615.

[12] Zhao J, Zhu H, Niu L. BiTNet: A lightweight object detection network for real-time classroom behavior recognition with transformer and bi-directional pyramid network[J]. Journal of King Saud University-Computer and Information Sciences, 2023, 35(8): 101670.

[13] Fang B, Li X, Han G, et al. Facial expression recognition in educational research from the perspective of machine learning: A systematic review[J]. IEEE Access, 2023, 11: 112060-112074.

[14] Trabelsi Z, Alnajjar F, Parambil M M A, et al. Real-time attention monitoring system for classroom: A deep learning approach for student's behavior recognition[J]. Big Data and Cognitive Computing, 2023, 7(1): 48.

[15] Dhara T, Singh P K, Mahmud M. A fuzzy ensemble-based deep learning model for EEG-based emotion recognition[J]. Cognitive Computation, 2024, 16(3): 1364-1378.

[16] Moise G, Dragomir E G, Șchiopu D, et al. Towards integrating automatic emotion recognition in education: A deep learning model based on 5 EEG channels[J]. International Journal of Computational Intelligence Systems, 2024, 17(1): 230.

[17] Ramadan M A, Salem N M, Mahmoud L N, et al. Multimodal machine learning approach for emotion recognition using physiological signals[J]. Biomedical Signal Processing and Control, 2024, 96: 106553.

[18] Yang X, Yan H, Zhang A, et al. Emotion recognition based on multimodal physiological signals using spiking feed-forward neural networks[J]. Biomedical Signal Processing and Control, 2024, 91: 105921.

[19] Guo Z, Yang M, Lin L, et al. E-MFNN: an emotion-multimodal fusion neural network framework for emotion recognition[J]. PeerJ Computer Science, 2024, 10: e1977.

[20] Geetha A V, Mala T, Priyanka D, et al. Multimodal emotion recognition with deep learning: advancements, challenges, and future directions[J]. Information Fusion, 2024, 105: 102218.

[21] Zhang S, Yang Y, Chen C, et al. Deep learning-based multimodal emotion recognition from audio, visual, and text modalities: A systematic review of recent advancements and future prospects[J]. Expert Systems with Applications, 2024, 237: 121692.

[22] Shiri F, Ahmadi E, Rezaee M, et al. Detection of student engagement in E-learning environments using EfficientNetV2-L together with RNN-based models[J]. Journal of Artificial Intelligence, 2024, 6: 85.