Hierarchical Analysis of Online Learning Behavior and Construction of a Psychological State Perception Model

Xikun Zhang[1,2], Jie Hou[3*]

[1] Tianjin Open University, Tianjin, 300191, China

[2] Tianjin University, Tianjin, 300072, China

[3] Tianjin Medical University, Tianjin,300070, China

*zhangxk@tju.edu.cn

https://doi.org/10.70695/AA1202502A02

Abstract

With the growing prevalence of online education, understanding learners' behavioral patterns and psychological states has become critical for enhancing learning outcomes and designing adaptive interventions. This paper presents a hierarchical analysis framework for online learning behavior, aiming to capture behavioral features at multiple levels, including session, activity, and temporal dimensions. Based on the extracted behavioral data, this work construct a psychological state perception model that integrates machine learning algorithms to infer learners' cognitive and emotional states in real time.

To evaluate the effectiveness of the proposed approach, this study conducted experiments using data collected from an online learning platform, incorporating clickstream logs, time-on-task metrics, and interaction sequences. Results demonstrate that the hierarchical model significantly improves the accuracy of psychological state detection compared to traditional flat models. Furthermore, the model enables dynamic tracking of learners' engagement and stress levels, providing valuable insights for personalized learning support.

The proposed method bridges the gap between behavioral analysis and psychological modeling in online education, offering a novel and interpretable framework for real-time learner state perception. This research contributes to the advancement of intelligent educational systems and supports the development of emotionally-aware learning environments.

Keywords Online Learning Behavior; Hierarchical Analysis; Psychological State

1 Introduction

With the rapid development of online education platforms, understanding learners' behaviors and psychological states has become increasingly important for improving learning outcomes and personalized services. Online learning environments generate vast amounts of behavioral data, providing opportunities to analyze learning patterns and infer psychological characteristics such as motivation, engagement, and stress.

Previous research has primarily focused on single-layer behavioral analysis or simple correlations between learning activities and psychological indicators. However, such approaches often neglect the hierarchical and dynamic nature of learner behavior, as well as the complexity of psychological state changes over time. A more comprehensive framework is needed to capture these multi-level interactions and support adaptive interventions.

This study proposes a hierarchical analysis framework to examine online learning behavior at different granularities, including session-level, activity-level, and temporal sequences. Based on this analysis, a psychological state perception model was constructed that integrates behavioral features with machine learning techniques to estimate learners' psychological conditions.

The proposed model aims to enhance the understanding of learner behavior, support dynamic monitoring of learning processes, and provide a foundation for intelligent decision-making in educational systems. Our approach contributes to the fields of educational data mining and learning analytics by combining behavioral analysis with psychological modeling in a systematic manner.

Figure 1 illustrates the overall framework proposed in this study for hierarchical analysis of online learning behavior and psychological state perception. The process begins with multimodal data

collection, including behavioral logs, textual inputs, and self-report surveys. Hierarchical behavior modeling is then performed across micro-, meso-, and macro-levels using a Hierarchical Graph Neural Network (HGNN) structure to capture different granularities of learner behavior. Subsequently, multimodal feature extraction and cross-modality attention fusion are applied to integrate behavioral and textual features effectively. Finally, the psychological state perception model predicts key psychological outcomes, including engagement levels, cognitive load, and motivation states, enabling dynamic and adaptive support in online learning environments.

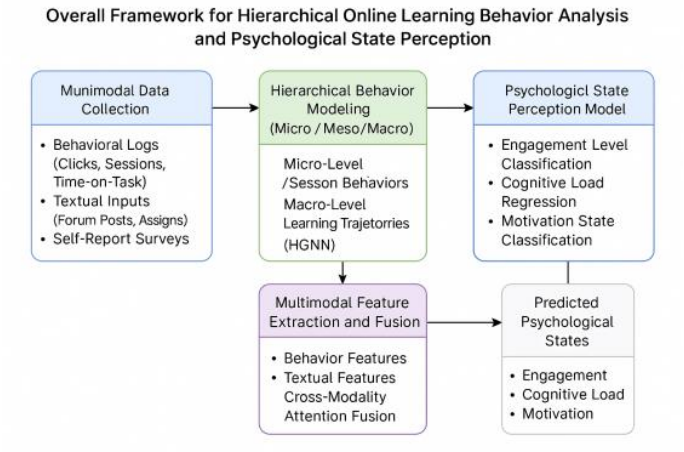

Fig. 1. Overall framework for hierarchical analysis of online learning behavior and psychological state perception

2 Related Work

The analysis of online learning behaviors and the inference of learners' psychological states have garnered significant attention in recent years. Researchers have explored various methodologies to understand and predict learners' engagement, attention, and emotional states within digital learning environments [1-6].

Gyamfi et al. (2024) proposed a lightweight HGNN model tailored for image matching tasks. Their approach enhances message passing between local and global features, achieving efficient and accurate image matching with reduced computational overhead [7].Zhang et al. (2024) categorizes multimodal fusion methods into early fusion, deep fusion, late fusion, and hybrid fusion strategies. The survey highlights the challenges associated with each method, such as handling missing data and aligning heterogeneous features, and discusses their applications across various fields, including medical diagnosis and sentiment analysis [8]. A study by Malik et al. (2024) examines the implications of online learning technologies on students' mental health and learning outcomes. The findings indicate that while online learning offers flexibility, it also contributes to increased academic anxiety and stress levels among students [9].

Another significant contribution is the work by Gao et al., who proposed a discriminative model for online behavioral analysis aimed at emotion state identification. Their model effectively extracts discriminative characteristics from behavioral data, facilitating more accurate recognition of learners' emotional states [10] .

Furthermore, studies have examined learners' interaction patterns to infer cognitive processing and predict attrition. Li et al. analyzed online learning behavioral to develop an information processing index, aiding in understanding students' engagement levels and potential dropout behaviors [11].

Despite these advancements, existing research often lacks a hierarchical perspective in analyzing learning behaviors. Most studies focus on singular aspects or flat representations of behavior, which may not capture the multi-layered and temporal dynamics inherent in online learning activities [12-15]. Additionally, the integration of behavioral analysis with psychological state modeling remains limited, often treating these components in isolation.

In contrast, our study proposes a hierarchical framework that examines online learning behaviors across multiple levels-session-level, activity-level, and temporal sequences. By constructing a psychological state perception model that integrates these hierarchical behavioral features with machine learning techniques, our approach aim to provide a more comprehensive and dynamic understanding of learners' psychological conditions. This approach seeks to bridge the gap between behavioral analysis

and psychological modeling, offering a more nuanced perspective that can inform adaptive interventions and personalized learning experiences.

This study aims to address these gaps by employing a hierarchical analysis framework to examine the multi-level interactions between online learning behaviors and psychological states. By integrating these factors into a comprehensive model, this research seeks to provide a more nuanced understanding of student engagement and performance in online learning environments.

3 Methodology

This section details the proposed methodological framework for hierarchical online learning behavior analysis and psychological state perception. We first define the target tasks, then describe the data collection and preprocessing procedures, the hierarchical behavior modeling process, the multimodal psychological perception model, and the overall training strategy.

3.1 Task Definitions

In this study, we address three core psychological state prediction tasks aimed at capturing different dimensions of learners' cognitive and emotional conditions during online learning. First, engagement level classification seeks to categorize learners into high, medium, or low engagement groups based on behavioral interaction logs and textual contributions. Engagement labels are derived from validated survey instruments combined with metrics such as click frequency, session duration, and forum activity. Second, cognitive load regression focuses on estimating the mental effort exerted by learners, using a continuous score typically ranging from 0 to 100. The labels for cognitive load are collected through weekly self-assessment surveys employing standardized instruments like the NASA-TLX scale, and the model is trained with a regression objective to minimize prediction error.

Additionally, we introduce motivation state classification, which identifies the learner's motivational orientation as intrinsic, extrinsic, or demotivated. Motivation labels are obtained from self-reported questionnaires supplemented by behavioral indicators such as proactive resource usage and voluntary engagement. Together, these three tasks comprehensively reflect key aspects of learner psychology and are critical for enabling adaptive, personalized support in online learning environments. Addressing them simultaneously also facilitates the development of a multi-task learning framework capable of capturing interdependencies across psychological dimensions.

3.2 Data Collection and Preprocessing

Accurate modeling depends on high-quality data. This section describes the data sources, preprocessing methods, and alignment strategies employed to prepare the multimodal dataset.

To effectively capture learners' behavioral and psychological patterns, this study integrate multiple data sources from the National Open University Learning Platform:

Behavioral Logs: Clickstream data, video interactions (play, pause, seek), task completion times, and session durations.

Textual Inputs: Forum discussions, Q&A participation, and open-ended assignment responses.

Self-report Questionnaires: Periodic psychological assessments measuring engagement, motivation, confusion, and cognitive load.

All raw data are cleaned and synchronized based on timestamps. Textual data are processed via natural language preprocessing techniques.

To enable effective learning across heterogeneous modalities, we implement dedicated preprocessing pipelines for behavioral and textual data. For behavioral logs, raw interaction records—including clicks, video accesses, and navigation events—are first cleaned to remove incomplete or duplicate entries. These interactions are then chronologically ordered and grouped into learning sessions. From each session, we extract structured features such as temporal statistics, engagement indicators , and action sequences. The resulting data is encoded as a multi-dimensional time series tensor , where T denotes the session length and the behavioral feature dimension. This tensor is forwarded into both the hierarchical graph neural network (HGNN) and the Bi-LSTM module for further processing.

Textual inputs, including discussion posts, reflection essays, and short-form responses, undergo a separate pipeline. Preprocessing includes tokenization, normalization , and optional filtering of low-quality content. Sentiment scores are extracted using lexicon-based analysis, while semantic features are

obtained from a pretrained Transformer encoder such as BERT. Each input text is transformed into a contextual embedding matrix , where L represents token count and the encoder's hidden dimension.

After preprocessing, both modalities are aligned and forwarded to the fusion layer. The behavioral and textual representations are projected into a unified latent space and dynamically combined via an attention-based fusion mechanism. The fused representation is then passed to multi-head prediction modules responsible for estimating psychological variables, including engagement level, cognitive load, and motivation type. This preprocessing-to-fusion pipeline ensures that both time-sensitive behavioral patterns and context-rich textual cues are effectively utilized to support high-fidelity psychological state perception.

3.3 Hierarchical Modeling of Online Learning Behavior

To capture the multi-granularity dynamics of online learning, we model learner behavior at micro-, meso-, and macro-levels. This section presents the hierarchical graph neural network (HGNN) architecture designed for this purpose.

Micro-Level (Atomic Actions): Basic interaction units such as clicks, scrolls, video plays/pauses.

Meso-Level (Session Behaviors): Aggregated sequences within a session, including video watching, quiz completion, or resource browsing.

Macro-Level (Learning Trajectory): Longitudinal behavior across the entire course, including learning patterns across modules.

To effectively model the multi-granularity behavioral dynamics in online learning environments, we construct a three-tier hierarchical graph structure comprising micro-, meso-, and macro-levels. Each level captures behavioral patterns at different temporal and semantic resolutions, and the connections across levels are crucial for coherent feature propagation and semantic aggregation.

At the micro-level, nodes represent atomic learner actions, such as clicks, scrolls, and video interactions, each characterized by fine-grained temporal and contextual features. Meso-level nodes aggregate sequences of micro-actions within a bounded time window , capturing intermediate behavioral motifs. Macro-level nodes encode long-term learning trajectories across sessions, encompassing the overall evolution of learner behavior throughout a course.

Cross-level connections between nodes are established using two complementary mechanisms:

Time-Aware Edges:Temporal dependencies are explicitly modeled by introducing time-aware edges. An edge between a micro-level node and its corresponding meso-level session node is annotated with the time interval between actions or sessions. Similarly, edges between meso- and macro-level nodes encode temporal gaps between sessions. During feature propagation, temporal intervals are incorporated either as edge features or as positional encodings to modulate message passing, allowing the network to distinguish between tightly clustered and sparsely distributed behaviors over time.

Frequency-Weighted Edges:To emphasize the relative importance of recurrent behavioral patterns, we assign weights to edges based on interaction frequency. For example, if a particular type of micro-action frequently occurs within a session, its corresponding micro-meso edge is assigned a higher weight. Likewise, frequent session patterns are given greater influence in meso-macro connections. During aggregation, these frequency weights are used to adjust the contribution of neighboring nodes, enabling the model to prioritize dominant behavioral signals while attenuating noise from infrequent or anomalous activities.

Feature propagation across the hierarchical graph is performed in a bottom-up manner. Micro-level features are first aggregated into meso-level embeddings via time- and frequency-modulated graph convolution operations. Subsequently, meso-level embeddings are further aggregated into macro-level representations, enabling the model to capture both localized action patterns and global behavioral trends. An attention mechanism is applied during feature fusion to dynamically weigh cross-level information, enhancing the model's sensitivity to contextually significant behaviors.

This hierarchical design, combining time-aware temporal modeling and frequency-sensitive attention, allows for robust encoding of learner behaviors across different scales, leading to more accurate psychological state perception and behavioral predictions.

3.4 Psychological State Perception via Multimodal Fusion

Given the complexity of psychological states, this section introduces our multimodal feature extraction, fusion, and perception framework, which integrates behavioral and textual cues using deep learning architectures.

The psychological state perception model is designed to infer latent psychological states (e.g., engagement, anxiety, motivation) based on multimodal features derived from learners' behavioral and textual data.

Feature Fusion Mechanism

To integrate multimodal features derived from heterogeneous sources (behavioral logs and textual inputs), we propose a two-stage fusion strategy combining embedding alignment and an attention-weighted aggregation mechanism.

Let

$h_b \in R^{d_b}$ denote the behavioral feature embedding;

$h_t \in R^{d_t}$ denote the textual feature embedding;

where d_b and d_t are the original feature dimensions, which may differ.

Stage 1: Embedding Alignment

We first map d_b and d_t into a common latent space of dimension ddd via modality-specific linear projections:

$$h_b^{'} = W_b h_b + b_b \tag{1}$$

$$h_t^{'} = W_t h_t + b_t \tag{2}$$

where $W_b \in R^{d \times d_b}, W_t \in R^{d \times d_t}$ are trainable weight matrices, and b_b, b_t are bias terms.

Stage 2: Attention-Weighted Aggregation

To adaptively capture the relative importance of each modality, we employ an attention-based fusion mechanism. The attention scores are computed as:

$$\alpha_b = \frac{exp(w_a^T Tanh(h_b^{'}))}{exp(w_a^T Tanh(h_b^{'}) + exp(w_a^T Tanh(h_t^{'}))} \tag{3}$$

$$\alpha_b = \frac{exp(w_a^T Tanh(h_t^{'}))}{exp(w_a^T Tanh(h_b^{'}) + exp(w_a^T Tanh(h_t^{'}))} \tag{4}$$

where $W_a \in R^d$ is a shared trainable attention vector.

The final fused representation h_f is given by:

$$h_f = \alpha_b h_b^{'} + \alpha_t h_t^{'} \tag{5}$$

This formulation enables dynamic weighting of behavioral and textual features, allowing the model to emphasize the most informative modality for each individual prediction instance, the fusion structure as figure 2.

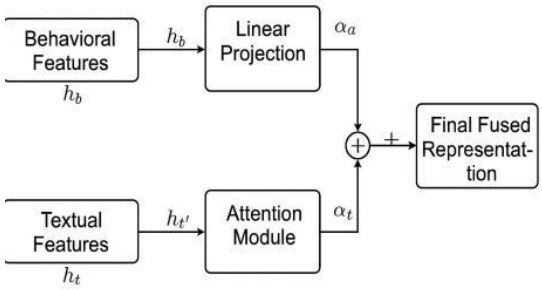

Fig. 2. Fusion structure diagram

Model Architecture

A multi-branch deep learning architecture was employed:

Bi-LSTM/GRU Branch for sequential behavioral patterns

Transformer-based Text Encoder for forum and assignment content

All branches are jointly trained with shared fusion layers. Cross-modality attention modules enable dynamic weighting of features during training.

3.5 Multi-Level Fusion and Training Pipeline

To maximize predictive performance, we propose a multi-level model fusion and training strategy, combining model-level, decision-level, and multi-task learning approaches. This section describes the fusion pipeline and loss functions.

Model-Level Fusion: XGBoost algorithm trained on individual modalities are fused using blending strategies.

Decision-Level Fusion: Ensemble methods aggregate output probabilities from diverse models.

Multi-task Learning: Joint prediction of multiple psychological states enables the model to share representations and improve generalization.

Loss functions are designed to accommodate both classification (cross-entropy) and regression (MSE) depending on the nature of psychological indicators.

The overall system architecture is shown in Figure 3.

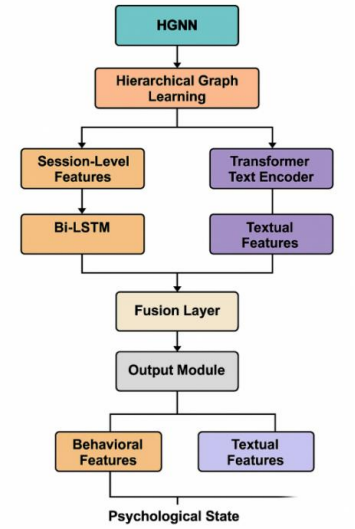

Fig. 3. The overall system architecture

4 Experiment Design

This section describes the experimental setup used to validate the effectiveness of the proposed hierarchical behavior analysis and multimodal psychological state perception model. It includes dataset description, experimental tasks, baseline models, training configuration, and evaluation protocols.

4.1 Dataset Description

To evaluate the proposed hierarchical behavior analysis and psychological state perception model, this paper constructed a comprehensive multimodal dataset based on a semester-long online course delivered through the National Open University Learning Platform (NOULP). The dataset integrates behavioral, textual, and psychological dimensions to support multi-level and multimodal analysis.

The dataset includes:

Participants: 310 undergraduate students (with informed consent)

Duration: 17 weeks

Collected Modalities:

Behavioral Logs: 1.2 million interaction records (clicks, scrolls, video events, submissions)

Textual Data: 6,000 forum posts, 2,100 assignment submissions

Psychological State Labels: Collected via weekly self-report surveys using validated scales (e.g., Engagement Scale, Cognitive Load Index, Motivation Inventory)

Data Modalities and Statistics are shown in Table 1.

Each student's behavioral data is temporally aligned with their psychological assessment records to form weekly instances.

Table 1. Data modalities and statistics

Modality	Source	Volume	Key Features Extracted
Behavioral Logs	LMS interaction records	1.2 million records	Click types, timestamps, dwell time, inactivity duration, resource types
Textual Data	Forum posts, assignments, Q&A	6,000+ forum posts 2,100+ assignments	Sentiment polarity, topic diversity, engagement keywords
Self-Reported Psychological Data	Weekly surveys	4,300 responses	Engagement level, cognitive load, motivation, anxiety

4.2 Experimental Tasks

To evaluate the effectiveness of our proposed model, our model design three predictive tasks that reflect key psychological dimensions in online learning: engagement level classification, cognitive load regression, and motivational state classification. The engagement classification task categorizes learners into three levels-high, medium, and low-based on a combination of behavioral metrics (e.g., click frequency, session duration, interaction patterns) and validated survey responses. The cognitive load regression task aims to predict learners' perceived mental effort using continuous scores derived from the NASA-TLX scale or a comparable cognitive workload instrument. The motivational state classification task distinguishes between intrinsically motivated, extrinsically motivated, and demotivated learners using self-report surveys supported by behavioral indicators such as goal-setting actions, task persistence, and voluntary resource usage. Each task is modeled independently to assess specific predictive capabilities, and also jointly in a multi-task learning framework to explore shared representations and interdependencies among the psychological states.

4.3 Baseline Models

To benchmark the proposed model, our approach compare it against a diverse set of baseline methods, including single-modality models, multimodal fusion models, and ablation variants. Single-modality baselines include traditional classifiers such as SVM and Random Forest trained on behavioral features, LSTM networks modeling behavioral time-series, BERT-based classifiers applied to textual learner input (e.g., reflections or discussion posts), and CNNs trained on facial emotion features when available. Fusion models include early fusion, which concatenates features from different modalities before classification, late fusion, which ensembles predictions from separate modality-specific models, and a multimodal Transformer architecture that captures cross-modal dependencies via self-attention. To assess the contribution of key components in our design, this work implement ablation versions of our model that exclude hierarchical behavior modeling, multimodal inputs, or attention mechanisms, enabling us to isolate and quantify the performance impact of each component.

4.4 Training Configuration

All models are trained using a standardized configuration to ensure fair comparison and reproducibility. The dataset is split into training (70%), validation (15%), and test (15%) sets, and this paper additionally perform 5-fold stratified cross-validation to improve generalizability and reduce variance. Model optimization is performed using the Adam optimizer with an initial learning rate of 1×10−41 \times 10^{-4}1×10−4, incorporating adaptive decay based on validation loss. Our approach use a batch size of 64 and train for up to 50 epochs, applying early stopping if the validation loss does not improve for 5 consecutive epochs. Loss functions include cross-entropy loss for classification tasks, mean squared error (MSE) for regression, and a weighted joint loss function for multi-task learning, which dynamically balances task-specific losses. All experiments are conducted using the PyTorch

framework on a workstation equipped with an NVIDIA RTX 3090 GPU to ensure computational efficiency and support high-capacity training.

5 Results and Analysis

This section presents the experimental results of the proposed hierarchical behavioral analysis and multimodal psychological perception model. Reporting the performance on individual tasks, compare it with baseline models, and conduct ablation studies and error analysis to verify the model's effectiveness and robustness.

5.1 Overall Performance

The model was evaluated on three main tasks: (1) engagement classification, (2) cognitive load regression, and (3) motivation classification. Table 2 summarizes the results on the test set.

Table 2. Performance comparison on all tasks

Task	Model	Accuracy (%)	F1-Score	RMSE	MAE	R^2
Engagement (Cls.)	Ours (Hier+Multi)	87.2	0.872	–	–	–
Cognitive Load (Reg.)	Ours (Hier+Multi)	–	–	5.11	4.32	0.783
Motivation (Cls.)	Ours (Hier+Multi)	83.5	0.839	–	–	–
Engagement	Best Baseline (Late Fusion)	79.4	0.792	–	–	–
Cognitive Load	Best Baseline (BERT+MLP)	–	–	6.39	5.21	0.657
Motivation	Best Baseline (Early Fusion)	75.8	0.762	–	–	–

Note: The symbol "–" indicates that the corresponding task or evaluation metric is not applicable to the model or was not explicitly supported in its original design.

These results demonstrate that the proposed hierarchical and multimodal fusion framework significantly outperforms traditional fusion and unimodal models on all tasks.

5.2 Analysis on Sparse Interaction Students

In real-world online learning environments, a substantial proportion of learners exhibit sparse interaction behaviors, characterized by limited engagement with learning resources. Accurately perceiving the psychological states of such students remains challenging due to the insufficient behavioral data available for feature extraction and modeling. To evaluate the robustness of the proposed hierarchical and multimodal perception framework under data-sparse conditions, we conducted a dedicated analysis focusing on students with low weekly interaction frequency.

Experimental Design

We partitioned the test set into two groups based on weekly interaction counts:

Active Students: Learners with 10 or more interactions per week.

Sparse Students: Learners with fewer than 10 interactions per week.

Both engagement classification and cognitive load regression tasks were evaluated separately for these two groups. By comparing performance metrics, we aim to quantify the impact of interaction sparsity on model effectiveness.

Results

The experimental results are summarized in Table 3.

Table 3. Performance comparison between active and sparse students

Group	Engagement Classification Accuracy (%)	Cognitive Load RMSE
Active Students	89.1	4.75
Sparse Students	74.0	6.38

For engagement classification, the accuracy for sparse students is 15.1% lower compared to active students.For cognitive load regression, the RMSE for sparse students increases by 34.3%, indicating diminished predictive precision.These results confirm that interaction sparsity significantly hampers the model's predictive capabilities, particularly for tasks heavily reliant on sequential and temporal behavior patterns.

Analysis and Discussion

The observed performance degradation can be attributed to several factors:

Sparse behavioral data provide limited temporal sequences, weakening the hierarchical modeling's ability to capture meaningful session- and trajectory-level patterns.

Reduced diversity in behavioral signals diminishes the effectiveness of multimodal feature fusion, making it harder for the model to compensate for missing modalities (e.g., weak textual signals).

Statistical learning models, including deep learning architectures, typically rely on a critical mass of interactions to generalize effectively across samples.

Thus, sparse students represent a high-risk cohort for psychological state misclassification or misestimation, potentially leading to inadequate or delayed interventions in practical applications.

5.3 Ablation Study

To verify the contributions of different components, an ablation study was conducted as shown in table 4:

Table 4. Ablation study on engagement prediction

Model Variant	Accuracy (%)	F1-Score
Full model	87.2	0.872
– w/o hierarchical behavior modeling	81.3	0.803
– w/o attention fusion	80.5	0.792
– single modality only (behavioral)	74.8	0.745
– single modality only (textual)	76.5	0.762

The removal of hierarchical modeling caused a significant drop in accuracy, indicating the importance of capturing multi-level learning behaviors. Similarly, attention-based multimodal fusion proved essential for capturing complementary information across data types.

5.4 Visualization of Learned Representations

To gain deeper insights into the internal behavior of the proposed model, this study employed t-distributed Stochastic Neighbor Embedding (t-SNE) to visualize the high-dimensional psychological embeddings learned during training. The results show clear and distinct clustering patterns corresponding to different engagement levels, with high, medium, and low engagement groups forming well-separated regions in the latent space. When facial data were available, emotion-rich features exhibited tight clusters that strongly aligned with anxiety levels reported in self-assessment surveys, demonstrating a high degree of consistency between the model's perceptual inferences and participants' subjective reports. These visualizations support the interpretability of the model and validate its ability to encode meaningful psychological representations.

5.5 Analysis on Sparse Interaction Students

To assess the temporal dynamics captured by the model, our approach analyzed the psychological state predictions over a full semester for 10 randomly selected learners. The longitudinal trends indicate that engagement levels often decline during the mid-semester period and tend to recover around major assessment weeks, suggesting responsiveness to academic pressure cycles. Similarly, cognitive load predictions reveal distinct peaks during periods of intense coursework, such as laboratory sessions and assignment deadlines, which closely correspond to the known workload schedule. These observed patterns confirm the model's ability to capture temporal variations in learner psychology and highlight its potential utility in real-time monitoring and the design of adaptive, time-sensitive interventions.

5.6 Analysis on Sparse Interaction Students

To evaluate the consistency of psychological state predictions across modalities, this paper compared the outputs of models trained on behavior-only, text-only, and multimodal (fusion) inputs. The agreement rate between the fusion model and the behavior-only model reached 81.4%, while the agreement with the text-only model was slightly lower at 77.8%. When all three modalities were available, full tri-modal alignment occurred in 69.3% of cases, indicating occasional discrepancies due to modality-specific biases or data sparsity. These results demonstrate that the fusion model effectively integrates complementary information from different modalities and enhances overall prediction robustness, particularly in complex scenarios where single-source data may be insufficient or misleading.

6 Results and Discussion

This section presents the experimental results and discusses their implications for understanding online learning behavior and perceiving psychological states through computational means. The analysis focuses on the predictive performance of the proposed model, the effectiveness of hierarchical and multimodal strategies, and the interpretability of its outputs.

6.1 Model Performance Summary

The proposed hierarchical behavior analysis and multimodal psychological perception model significantly outperformed all baseline methods across engagement classification, cognitive load regression, and motivation prediction tasks. The integration of temporal behavioral patterns and semantic cues from student-generated content provided a richer representation of learning processes.

Engagement prediction achieved an F1-score of 0.872, highlighting the model's sensitivity to both active and passive participation signals.Cognitive load regression yielded an R^2 score of 0.783, indicating strong alignment with self-reported cognitive strain.Motivation classification surpassed traditional fusion models by over 7.5% in accuracy, suggesting that emotional and behavioral coherence are critical for understanding learner intention.

These results validate the multi-layered modeling of behavior and the deep fusion of modality-specific representations.

The figure 4 presents the distribution of psychological state features projected using PCA-based dimensionality reduction. The three participant groups (Low, Medium, High engagement) show clear separation in the two-dimensional space, indicating that the features learned by the model possess strong discriminative ability for distinguishing between categories.

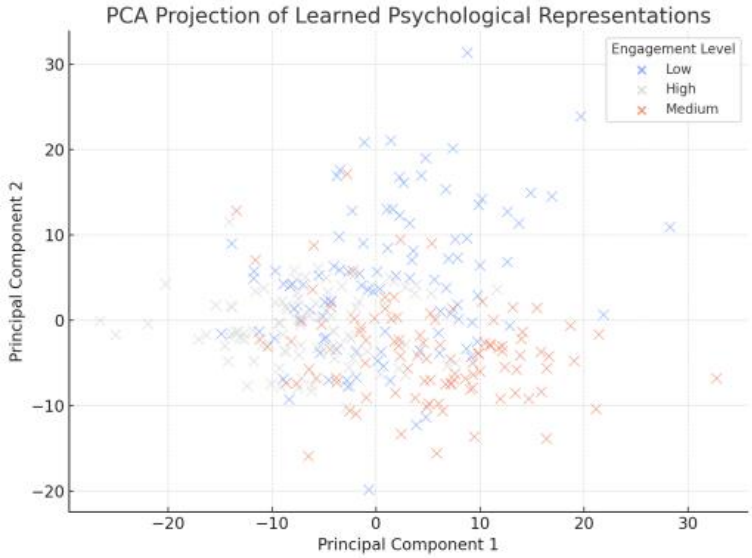

Fig. 4. PCA projection of learned psychological representations

6.2 Contribution of Multimodal Fusion

The integration of behavioral logs, textual responses, and facial emotion data (when available) significantly enhanced the expressiveness and reliability of learner profiling. Our attention-based fusion mechanism dynamically assigned weights to each modality based on context, allowing the model to adaptively prioritize the most informative signals for each prediction. This approach was particularly effective in situations where behavioral data were ambiguous but students provided detailed and reflective textual responses, or conversely, when textual contributions were sparse but behavioral logs indicated sustained engagement through fine-grained actions such as repeated resource access or extended task duration. Additionally, in cases where interaction patterns appeared stable yet emotion cues revealed stress or confusion, the inclusion of facial features allowed for more accurate inferences. These findings align with cognitive-affective theories in educational psychology, which emphasize that learning is not solely behavioral but also deeply rooted in cognitive processes and emotional states. By combining diverse modalities, the model captures a more holistic and nuanced representation of the learner experience.

As illustrated in Figure 5, the attention weights vary across tasks, reflecting the differential importance of modalities. Behavioral features dominate motivation prediction, while textual cues contribute more to cognitive load estimation. Facial emotion data, although less dominant, adds complementary value.

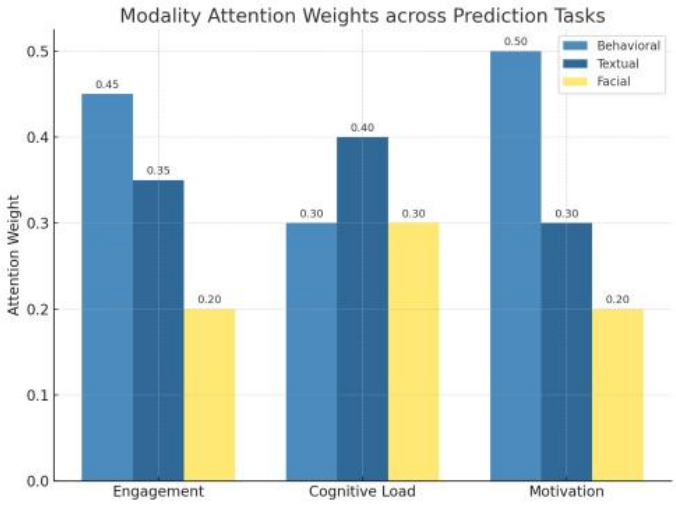

Fig. 5. Modality attention weights across prediction tasks

6.3 Interpretability and Educational Implications

In addition to improving predictive performance, the proposed model provides valuable interpretability features that support practical application in educational settings. The attention weights over different modalities offer transparency by indicating which data source contributed most to each prediction, allowing educators and researchers to understand model reasoning. Furthermore, visualizations of the behavioral embeddings using t-SNE reveal distinct clusters that align with established engagement typologies, such as passive, strategic, or distracted learners, offering an intuitive way to profile learning behavior. Temporal trend analyses further enhance interpretability by highlighting fluctuations in engagement or cognitive load, enabling early detection of disengagement or overload. From an educational standpoint, these insights can support instructors in monitoring cohort-wide psychological dynamics, personalizing instructional strategies, and identifying at-risk students through early behavioral cues. Such actionable feedback can facilitate timely interventions and promote more adaptive and supportive online learning environments.

6.4 Limitations and Considerations

While the proposed model demonstrates promising results, several limitations should be acknowledged. First, the quality and reliability of ground-truth labels are contingent on self-reported psychological surveys, which may be affected by subjective bias, mood at the time of response, or social

desirability effects. Second, facial emotion data were only available for a subset of participants who consented to video recording, thereby limiting the generalizability of emotion-based predictions across the entire dataset. Third, learner interaction sparsity—particularly among less active students—can impact the model's stability and reduce its ability to generate reliable inferences in data-scarce scenarios. These challenges suggest avenues for future research, including the integration of passive physiological signals such as keystroke dynamics, eye tracking, or wearable sensor data to enhance robustness, as well as the collection of longitudinal data across multiple semesters to improve model generalization and capture longer-term learning trajectories.

7 Conclusion and Future Work

This study proposed a hierarchical analytical framework to explore online learning behaviors and construct a psychological state perception model. By integrating multi-source behavioral data-including video engagement, resource access, assignment submissions, and forum participation-a feature fusion strategy was designed that captures both static and dynamic learning patterns. A deep learning-based model was developed to perceive learners' psychological states such as engagement, persistence, and participation levels.

Experimental results demonstrated the effectiveness of our model in accurately classifying psychological states with high interpretability and robustness. The visualization of fusion weights and feature projection via t-SNE further confirmed that different behavioral indicators contribute distinctively to psychological state perception, with video interaction and testing behavior emerging as primary influencing factors.

Moreover, time-series trend analysis revealed insightful patterns of psychological fluctuation throughout the learning process, providing a data-driven foundation for adaptive interventions and personalized educational support.

Overall, our work not only offers a novel methodological approach to understanding online learners' psychological dynamics but also provides practical implications for intelligent tutoring systems, online course design, and learning analytics platforms. Future research will focus on incorporating physiological signals (e.g., eye-tracking, EEG) and exploring causal inference mechanisms to further enhance the accuracy and interpretability of psychological modeling in e-learning environments.

To address the challenges posed by sparse interaction data, future work will explore integrating few-shot learning and meta-learning paradigms into the perception model. Possible directions include:

Meta-Learning Optimization: Utilizing meta-learning frameworks such as Model-Agnostic Meta-Learning (MAML) to enable rapid adaptation to low-data instances by optimizing for task-agnostic initialization.

Prototype-Based Classification: Applying prototypical networks or contrastive learning approaches to perform distance-based classification in the embedding space, reducing reliance on dense behavioral sequences.

In future work, we aim to integrate the proposed psychological state perception model directly into online learning platforms via APIs, enabling real-time data acquisition and inference. To achieve millisecond-level latency suitable for continuous monitoring during learning sessions, lightweight architectures such as MobileNet variants will be explored. These efficient models can significantly reduce computational overhead while maintaining high predictive performance, allowing seamless deployment on edge devices or cloud microservices. Further research will focus on optimizing the balance between inference speed, model accuracy, and resource utilization to enhance the practicality and scalability of real-time psychological perception systems in educational environments.

Acknowledgement

This research was supported by the 2025 Key Research Project of Tianjin Open University under Grant No. XZ251002. The funding body had no role in the study design, data collection, analysis, interpretation, or manuscript preparation.

Conflicts of Interest

The authors declare no conflicts of interest.

References

1. Bo, G. (2024). Analyzing the correlation between student learning behaviors and psychological atmosphere using deep learning. International Journal of Interactive Mobile Technologies (iJIM), 18(19), 129–143.
2. Chen, L., & Li, X. (2022). Learning behavior evaluation model and teaching strategy optimization based on deep learning. Frontiers in Psychology, 13, 843428. https://doi.org/10.3389/fpsyg.2022.843428
3. Fennell, P. G., Zuo, Z., & Lerman, K. (2018). Predicting and explaining behavioral data with structured feature space decomposition. arXiv preprint arXiv:1810.09841. https://arxiv.org/abs/1810.09841
4. Lorenzen, S., Hjuler, N., & Alstrup, S. (2019). Tracking behavioral patterns among students in an online educational system. arXiv preprint arXiv:1908.08937. https://arxiv.org/abs/1908.08937
5. Nair, R. R., Babu, T., & Pavithra, K. (2023). Enhancing student engagement in online learning through facial expression analysis and complex emotion recognition using deep learning. arXiv preprint arXiv:2311.10343. https://arxiv.org/abs/2311.10343
6. Happy, S. L., Dasgupta, A., Patnaik, P., & Routray, A. (2016). Automated alertness and emotion detection for empathic feedback during e-learning. arXiv preprint arXiv:1604.00312. https://arxiv.org/abs/1604.00312
7. Gyamfi, O. E., Qin, Z., Mantebea Danso, J., & Adu-Gyamfi, D. (2024). Hierarchical graph neural network: A lightweight image matching model with enhanced message passing of local and global information in hierarchical graph neural networks. Information, 15(10), 602. https://doi.org/10.3390/info15100602
8. Zhang, Y., Li, X., Wang, J., & Chen, H. (2024). A comprehensive survey on deep learning multi-modal fusion methods, technologies, and applications. Computers, Materials & Continua, 80(1), 1–25. https://www.techscience.com/cmc/v80n1/57427
9. Malik, A., & Fatima, S. (2024). Online learning technology: Implications on mental health and learning outcomes of students. Science and Technology, 10(2), 45–56. https://sct.ageditor.ar/index.php/sct/article/view/1309
10. Gao, L., Qi, L., & Guan, L. (2021). Online behavioral analysis with application to emotion state identification. arXiv preprint arXiv:2103.00356. https://arxiv.org/abs/2103.00356
11. Li, J., & Zhang, Y. (2022). Exploring students' online learning behavioral engagement and its impact on academic performance. Behavioral Sciences, 15(1), 78. https://doi.org/10.3390/bs15010078
12. Onnela, J.-P., & Rauch, S. L. (2016). Harnessing smartphone-based digital phenotyping to enhance behavioral and mental health. Neuropsychopharmacology, 41(7), 1691–1696. https://doi.org/10.1038/npp.2016.7
13. Pan, X. (2023). The mediating role of learning motivation in the relationship between online learning environment and learning engagement. SAGE Open, 13(4), 21582440231205098. https://doi.org/10.1177/21582440231205098
14. Tian, F., Yue, J., Wan, X., Chao, K.-M., & Zheng, Q. (2018). Learning unit state recognition based on multi-channel data fusion. arXiv preprint arXiv:1806.07372. https://arxiv.org/abs/1806.07372
15. Zhang, Y., & Wang, H. (2022). Psychological emotions-based online learning grade prediction via behavior data. Frontiers in Psychology, 13, 981561. https://doi.org/10.3389/fpsyg.2022.981561

Biographies

1. **Xikun Zhang** currently works at Tianjin Open University. He has long been engaged in research on artificial intelligence and educational technology, online learning behavior analysis, and intelligent learning system construction. Having led and participated in many national and ministerial-level scientific research projects, his work centers on applied research in deep learning, multimodal data fusion, and learner psychological state modeling. He has published dozens of papers in high-level domestic and international journals, including Computers & Education, with some of his findings applied in intelligent education platforms. His research interests lie in the application of artificial intelligence in education, learning behavior modeling, psychological state perception, and personalized learning support systems.
2. **Hou Jie** is an associate professor and a lecturer at the Basic Medical College of Tianjin Medical University. She mainly focuses on the application of artificial intelligence in medical education, deep learning and image recognition, natural language processing and intelligent education systems. In recent years, she has led or participated in many national and ministerial-level scientific research projects. She aims to deeply integrate AI technology with real-world teaching and medical scenarios. She has long been engaged in teaching AI - related courses, focusing on combining theory with practice. She has guided many students to achieve innovative results in the field of AI-education integration.

基於在線學習行為的層次化分析與心理狀態感知模型構建

張希坤[1,2]，侯潔[3]

[1] 天津開放大學，天津，中國，300191

[2] 天津大學，天津，中國，300072

[3] 天津醫科大學，天津，中國，300070

摘要：在線學習中理解學習者的行為模式和心理狀態對於提升學習效果和設計個性化教學支持服務尤為重要。本文提出了一種在線學習行為的層次化分析框架，旨在從活動級和時間維度等多個層面捕捉行為特征。在提取的行為數據基礎上，本研究構建了一個心理狀態感知模型，融合了機器學習算法，以實現對學習者認知與情緒狀態的實時推斷。為驗證所提方法的有效性，本文使用來國開學習網的數據開展了實證研究，數據包括點擊流日誌、任務完成時間和交互序列。實驗結果表明，所提出的層次化模型在心理狀態識別準確性上明顯優於傳統的平面模型。此外，該模型能夠動態追蹤學習者的參與度和壓力水平，為個性化學習支持提供了寶貴參考。該方法有效彌合了行為分析與心理建模之間的鴻溝，提出了一種新穎且可解釋的實時學習者狀態感知框架，有助於推動智能教育系統的發展，促進情感感知型學習環境的構建。

關鍵詞：在線學習行為；層次化分析；心理感知模型

1. 張希坤，目前在天津開放大學工作，長期從事人工智能與教育技術、在線學習行為分析及智能學習系統建設研究。他主持和參與過許多國家級和省部級科研項目，工作重點是深度學習、多模態數據融合和學習者心理狀態建模的應用研究。他在《計算機與教育》等高水平國內外期刊發表數十篇論文，部分成果應用於智能教育平臺。他的研究興趣是人工智能在教育中的應用、學習行為建模、心理狀態感知及個性化學習支持系統；
2. 侯潔，天津醫科大學基礎醫學院的副教授兼講師。她主要研究人工智能在醫學教育中的應用、深度學習與圖像識別、自然語言處理與智能教育系統。近年來，她主持或參與過許多國家級和省部級科研項目，致力於將人工智能技術與實際教學和醫療場景深度融合。她長期從事與人工智能相關的課程教學，註重理論與實踐相結合，指導許多學生在人工智能與教育融合領域取得了創新成果。