IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Traitement du Signal
Vol. 42, No. 6, December, 2025, pp. 3135-3147
Journal homepage: http://iieta.org/journals/ts

A Multimodal Data Fusion Model for Personalized English Learning Path Recommendation Based on Image Processing

Qiaozhi Wen[1], Guoxi Chai[1*], Min Zhang[2]

[1] School of Foreign Languages, Yan'an University, Yan'an 716000, China
[2] School of Information Engineering, Shaanxi Xueqian Normal University, Xi'an 710010, China

Corresponding Author Email: chweboo@163.com

https://doi.org/10.18280/ts.420609

Received: 29 April 2025
Revised: 18 September 2025
Accepted: 10 October 2025
Available online: 31 December 2025

Keywords:
educational multimodal fusion, personalized learning path, cross-modal attention, feature decoupling, English teaching analytics

ABSTRACT

The rapid accumulation of multimodal data in English teaching environments presents new opportunities for personalized learning path recommendation. However, existing approaches struggle to effectively model the complex relationships between standardized instructional materials and non-standardized learner behaviors, resulting in insufficient precision in multimodal feature fusion and suboptimal adaptability of recommended learning paths. To address this limitation, a multimodal attention-driven teaching data fusion and personalized English learning path recommendation model (MATDF-ELR) was proposed. In consideration of the contextual specificity of English education, multimodal features are decoupled into teaching-dependent features, which represent cross-modal shared instructional content, and learning-diversity features, which capture individual variations in students' learning processes. The core innovation of the proposed framework lies in the teaching data dependency-diversity fusion module (TD3FM). Feature decoupling is guided through a multi-task loss function, while an interpretable cross-modal attention mechanism is employed to enable end-to-end mapping from multimodal teaching data to personalized learning paths. To evaluate the effectiveness of the proposed model, a multimodal English teaching dataset was constructed, comprising four-modal data collected from 120 students, along with expert-annotated learning paths. Experimental results demonstrated that MATDF-ELR achieved an F1@3 score of 0.87 in the learning path recommendation task, representing a 12.3% improvement over the strongest baseline method. In addition, the mutual information (MI) metric for multimodal feature fusion was improved by 18.2%. Visualization analyses further confirmed that the model is capable of accurately attending to pedagogically salient regions and anomalous learning behaviors. These findings establish a theoretically grounded and education-oriented technical framework for multimodal data mining in educational contexts and provide effective support for the intelligent deployment of personalized English instruction.

1. INTRODUCTION

The ongoing digital transformation has driven English education into a new phase characterized by multimodal data-driven instruction. Instructional carriers such as textbook text, blackboard images, spoken audio, and classroom video not only convey standardized pedagogical knowledge but also embed rich information regarding students' learning states and cognitive characteristics [1-3]. As a core component of adaptive education, personalized learning path recommendation aims to generate learning sequences that align with individual learner needs by modeling the interaction between teaching conditions and learning states [4, 5]. However, in current practice, the potential value of multimodal data has not yet been fully exploited, and substantial room for improvement remains in aligning recommended learning paths with students' actual cognitive demands.

Multimodal data processing in English teaching scenarios faces three fundamental challenges. First, pronounced semantic heterogeneity exists within instructional multimodal data: standardized teaching materials convey normative knowledge structures, whereas learning performance data reflect individualized cognitive deviations. This intrinsic disparity renders conventional cross-modal feature alignment methods ineffective in establishing meaningful associations between teaching and learning representations [6, 7]. Second, existing multimodal fusion approaches often integrate heterogeneous features in an undifferentiated manner, failing to distinguish cross-modal shared knowledge features from modality-complementary indicators of learning difficulty. As a result, the fused representations lack pedagogical specificity and are unable to adequately support precise learning path recommendations [8, 9]. Third, general-purpose multimodal models typically omit the cognitive principles inherent to educational contexts. Learning patterns such as knowledge transfer and dynamic variations in cognitive load are rarely modeled, leading to learning paths that deviate from foundational principles of learning science [10-12].

Notable limitations persist in existing studies. At the level of multimodal fusion, most approaches rely on shallow strategies such as simple concatenation or weighted summation, without designing fusion mechanisms tailored to the contextual specificity of instructional scenarios [13, 14]. Although generic cross-modal models have demonstrated strong performance on large-scale datasets, they remain poorly suited to the alignment requirements of English education, particularly the need to reconcile correct instructional exemplars with erroneous learner behaviors [15, 16]. At the level of learning path recommendation, prevailing methods primarily construct learner models using single-dimensional indicators such as test scores or assignment results, while neglecting the fine-grained cognitive state signals embedded in multimodal data. This limitation results in insufficient adaptability in recommended learning paths [17-20].

To address the aforementioned challenges, the MATDF-ELR was proposed. The model is grounded in the central hypothesis of educational multimodal feature decoupling, under which multimodal representations are separated into a teaching-norm-dependent subspace that characterizes cross-modal knowledge sharing and a learning-diversity subspace that reflects individual cognitive variation. Through the TD3FM, feature decoupling and targeted fusion are jointly guided by a multi-task loss function, thereby enabling an end-to-end mapping from raw multimodal data to interpretable personalized learning paths.

The main contributions can be summarized below. First, at the theoretical level, an educational multimodal feature decoupling hypothesis is formulated, and a corresponding mathematical modeling framework is established, providing a new theoretical perspective for feature representation in educational multimodal data. Second, at the methodological level, the TD3FM is designed, in which dependency loss and diversity loss are introduced to achieve feature separation and fusion under educational constraints, thereby enhancing the specificity and interpretability of multimodal representations. Third, at the resource level, the first multimodal dataset dedicated to English teaching, EMTD-2025, is constructed and released. The dataset comprises four-modal data collected from 120 students, accompanied by fine-grained cognitive state annotations, addressing a critical gap in publicly available data resources for this domain. Fourth, at the empirical level, comprehensive validation is conducted through quantitative experiments, ablation studies, and visualization analyses, collectively demonstrating the superiority of the proposed model in both multimodal fusion quality and learning path recommendation performance and providing robust support for practical deployment.

2. Methodology

2.1 Problem formulation

In instructional settings, the multimodal dataset is formalized as a four-tuple $D=\{T,I,A,V\}$, where T denotes the text modality, encompassing textbook knowledge-point texts and student assignment texts; I denotes the image modality, including photographs of classroom blackboard instruction and illustrative figures from teaching materials; A denotes the audio modality, which records speech data from students' oral practice; and V denotes the video modality, consisting of classroom interaction recordings that reflect students' cognitive states. According to the educational attributes of the data, the multimodal dataset is partitioned into two branches. Branch $M_A=\{T,I\}$ corresponds to static instructional materials that convey standardized knowledge content, whereas branch $M_B=\{A,V\}$ corresponds to dynamic learning performance, reflecting students' individualized cognitive states and learning behaviors. This grouping strategy establishes the structural foundation for subsequent feature decoupling, enabling the model to separately capture normative instructional information and personalized learning signals.

Let A denote a predefined set of English learning activities, including knowledge review, oral practice, and assignment error correction, which are aligned with instructional objectives. Given the multimodal data sequence Ds of a student s, the task of personalized learning path recommendation is to generate a learning activity sequence $P_s=[a_1,a_2,...,a_K]$, with $a_i \in A$, such that the sequence is adapted to the student's current cognitive state, progressively compensates for knowledge gaps, and optimizes learning efficiency. The rationality of the generated sequence is jointly constrained by the cognitive logic among learning activities and the student's real-time learning progress.

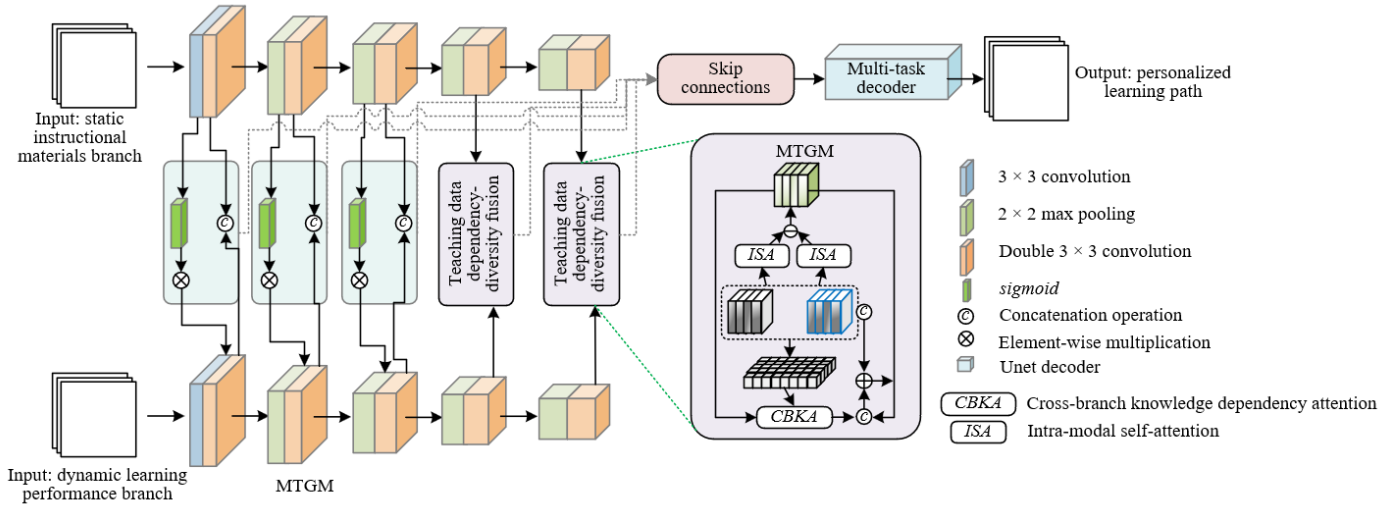

Input: static instructional materials branch
Input: dynamic learning performance branch
MTGM
Teaching data dependency-diversity fusion
Teaching data dependency-diversity fusion
Skip connections
Multi-task decoder
Output: personalized learning path
MTGM
ISA
ISA
CBKA
3 × 3 convolution
2 × 2 max pooling
Double 3 × 3 convolution
sigmoid
Concatenation operation
Element-wise multiplication
Unet decoder
CBKA Cross-branch knowledge dependency attention
ISA Intra-modal self-attention

Figure 1. Overall architecture of the MATDF-ELR model

2.2 Overall architecture of MATDF-ELR

The MATDF-ELR model is constructed upon a dual-branch encoder-decoder framework, which is explicitly designed to accommodate the grouping characteristics of educational multimodal data. Following a core logic of "separation-fusion-mapping," the architecture comprises four synergistic components.

First, dual-branch multimodal encoders are employed to process the static instructional materials branch M_A and the dynamic learning performance branch M_B, respectively. Partial convolutional kernel parameters are shared between the two branches, thereby constraining the model to learn cross-branch knowledge-common features while preserving modality-specific representations within each branch. Second, the initial three layers of the encoders are augmented with a Multimodal Teaching Guidance Module (MTGM), in which feature modulation is applied to suppress redundant background information in branch M_B and to enhance effective learning signals that are aligned with the knowledge content of branch M_A. Third, the final two encoder layers are replaced by the TD3FM, which constitutes the core innovation of the proposed architecture. Within this module, cross-branch knowledge dependency relationships and inter-modal learning diversity features are explicitly modeled, enabling precise fusion at a high semantic level. Finally, a multi-task decoder is employed to restore feature resolution through deconvolution operations while simultaneously optimizing two objectives: (i) the generation of learning path sequences, and (ii) the enforcement of consistency constraints for feature decoupling. Through this joint optimization, personalized learning paths that are well aligned with students' cognitive states are produced. An overview of the overall architecture is illustrated in Figure 1.

2.3 Dual-branch multimodal encoders

The dual-branch multimodal encoders are designed in response to the intrinsic differences between static instructional materials and dynamic learning performance. Differentiated feature extraction pipelines are constructed to capture education-relevant information through modality-specific processing and cross-modal interaction, while a combination of parameter sharing and separation strategies is adopted to balance general representation learning and domain-specific feature modeling. This design establishes a solid foundation for subsequent feature decoupling and precise fusion.

Branch A, which targets static instructional materials composed of text and images, follows a processing paradigm of modality-specific extraction followed by early cross-modal interaction, thereby ensuring consistent representation of standardized knowledge. For the text modality, inputs consist of textbook knowledge-point texts and assignment text sequences. Semantic features are extracted using an education-domain-pretrained Bidirectional Encoder Representations from Transformers (BERT) variant, which is fine-tuned on primary and secondary school English teaching corpora to enhance the capturing of core instructional elements such as grammatical rules and lexical collocations. The resulting textual embedding is expressed as:

$$E_t = BERT_{edu}(T) \tag{1}$$

In the equation, $\boldsymbol{E}_t \in \mathrm{R}^{L_t \times d_{emb}}$, where L_t denotes the text sequence length, and d_{emb} represents the embedding dimension. The image modality includes photographs of classroom blackboard writing and illustrative figures from teaching materials. These inputs are processed by an Edu-CNN optimized for instructional scenarios. Customized 3 × 3 convolutional kernels are employed to strengthen feature responses in textual regions and knowledge-point illustrations, while adaptive pooling is applied to preserve salient visual information. The resulting visual feature representation is given by:

$$F_i = Edu\text{-}CNN(I) \tag{2}$$

In the equation, $\boldsymbol{F}_i \in \mathrm{R}^{H_i \times W_i \times d_{conv}}$, with H_i and W_i denoting the spatial dimensions of the feature map and d_{conv} indicating the number of convolutional channels. To enable early alignment between abstract textual knowledge and visual information, both modalities are linearly projected into a unified feature space of dimension d_{fusion}. Subsequently, a cross-modal attention mechanism is employed to compute relevance weights between textual and visual features, guiding visual representations to attend to text-anchored knowledge points. The fused representation for Branch A is finally obtained as:

$$F_A = Cross\text{-}Attn(E_t, F_i) \tag{3}$$

This process effectively associates grammatical points and lexical items in the textual modality with their corresponding visual representations, thereby enhancing the semantic consistency and completeness of feature representations derived from static instructional materials.

Branch B focuses on dynamic learning performance data composed of audio and video modalities, with emphasis placed on extracting education-relevant features that reflect learners' cognitive states and their temporal evolution. For the audio modality, raw speech signals are transformed into Mel spectrograms S_a through pre-emphasis, framing, and windowing operations. The resulting representations are then fed into an Audio-CNN to extract acoustic features encompassing key indicators such as prosodic rhythm, oral fluency, and pronunciation accuracy. The output is expressed as:

$$F_a = Audio\text{-}CNN(S_a) \tag{4}$$

In the equation, $F_a \in \mathrm{R}^{T_a \times d_a}$, where T_a denotes the temporal length, and d_a represents the dimensionality of acoustic features. For the video modality, inputs consist of continuous classroom video frame sequences V_{seq}. A 3D-CNN model is employed to capture spatiotemporal features, in which dimensionality reduction is first performed using 1 × 1 × 1 convolutions, followed by 3 × 3 × 3 spatiotemporal convolutions to extract salient information. Visual cues such as facial expressions and gestural movements are explicitly encoded to quantify students' learning engagement levels. The output is expressed as:

$$F_v = 3D\text{-}CNN(V_{seq}) \tag{5}$$

In the equation, $F_v \in \mathrm{R}^{T_v \times d_v}$, with T_v denoting the frame sequence length and d_v indicating the visual feature dimension. Considering the temporal asynchrony between audio and video streams, linear interpolation is applied to align both feature sequences to a unified temporal resolution T. The aligned features are concatenated to form a joint representation $\mathit{Concat}(\boldsymbol{F}_a,\boldsymbol{F}_v)\in\mathrm{R}^{T\times(d_a+d_v)}$. This joint feature sequence is subsequently input into a Long Short-Term Memory (LSTM) network to model temporal learning state dynamics, capturing fluctuations in oral expression fluency and variations in classroom attentional engagement. The final integrated representation for Branch B is obtained as:

$$\boldsymbol{F}_B = \mathit{LSTM}(\mathit{Concat}(\boldsymbol{F}_a,\boldsymbol{F}_v)) \tag{6}$$

To optimize representational capacity and parameter efficiency, a "shared lower layers-independent higher layers" parameter configuration strategy is adopted in the encoder. Specifically, the base visual feature extraction layers, comprising the first three layers of the Edu-CNN and 3D-CNN, share convolutional kernel parameters, denoted as $\boldsymbol{W}^{v}_{share}$, while the base textual embedding layers share a common word embedding matrix $\mathbf{W}^{t}_{\mathrm{sh}are}$. This design is intended to uncover cross-branch generic feature patterns and reduce parameter redundancy and can be expressed as:

$$\boldsymbol{F}_{share} = \mathit{Shared\text{-}Params}(\boldsymbol{X}) \tag{7}$$

where, $\boldsymbol{X}$ represents the original visual or textual input. Through this strategy, the total number of model parameters is reduced by approximately 32%, while low-level feature associations between static instructional materials and dynamic learning performance are simultaneously strengthened. In contrast, the higher-level feature extraction layers are designed with branch-specific parameters. For Branch A, the high-level network employs instructional-norm-aware parameters, denoted as *Task-A-Params*, to focus on standardized characteristics such as knowledge-point structures and hierarchical relationships among instructional concepts, yielding:

$$F_A^{high} = \mathit{Task\text{-}A\text{-}Params}(F_{share}) \tag{8}$$

For Branch B, the high-level network utilizes learning-performance-aware parameters, denoted as *Task-B-Params*, to emphasize individualized features, including error patterns, ability deficiencies, and state fluctuations in learners' performance, producing:

$$F_B^{high} = \mathit{Task\text{-}B\text{-}Params}(F_{share}) \tag{9}$$

Through this strategy, model complexity and overfitting risk are effectively reduced via parameter sharing, while domain-specific feature extraction for the two data branches is preserved through independent parameterization. As a result, an organic balance between general representations and specialized representations is achieved.

2.4 MTGM

The MTGM is deployed in the first three encoder layers. Its primary objective is to leverage the normative characteristics of static instructional materials to guide selective attention over dynamic learning performance features, thereby suppressing redundant information unrelated to core instructional content and strengthening the representation of cross-modal relevant features. Standardized knowledge encoded in Branch A is fully utilized as a guiding signal to precisely modulate features in Branch B, enabling learning performance representations to focus more effectively on information aligned with instructional objectives and establishing a high-quality foundation for subsequent high-level feature fusion.

The operational mechanism of MTGM follows a three-stage process of feature association, weight generation, and feature modulation. First, the Branch A features $\boldsymbol{F}_A^{(l)}$ and Branch B features $\boldsymbol{F}_B^{(l)}$ at the *l*-th encoder layer are concatenated along the channel dimension, integrating normative instructional information with initial learning performance representations to form a cross-modal joint feature. Next, a lightweight Multilayer Perceptron (MLP) is applied to compress the feature dimension and to learn association patterns between the two branches. Through a sigmoid activation function, attention weights $W^{(l)}$ within the range [0,1] are generated, quantifying the relevance of each Branch B feature channel to the instructional content:

$$W^{l} = \sigma\left(MLP\left([F_A^{(l)};F_B^{(l)}]\right)\right) \tag{10}$$

where, [;] denotes channel-wise concatenation, σ represents the sigmoid activation function, and $W^{(l)}$ has the same dimensionality as $F_B^{(l)}$. Finally, the attention weights are applied to the original Branch B features via element-wise multiplication, achieving feature modulation:

$$F_B^{(l)} = F_B^{(l)} \odot W^{(l)} \tag{11}$$

In the modulated feature representation $F_B^{(l)*}$, components that are highly correlated with instructional content are amplified, while irrelevant or redundant information is suppressed. Through this process, MTGM establishes an early-stage alignment between instructional norms and learning performance, enhancing feature specificity and effectiveness, and providing support for the subsequent precise separation and fusion of dependency and diversity features.

2.5 TD3FM

The TD3FM constitutes the core innovative component of the proposed framework and is deployed in the final two encoder layers as well as at the decoder skip connections. Its primary objective is to simultaneously capture cross-modal shared instructional norms and student-specific learning variation features through feature decoupling and precise fusion, thereby providing fine-grained and highly targeted representations for personalized learning path recommendation. Unlike conventional multimodal fusion approaches, TD3FM introduces orthogonal subspace modeling guided by multi-task loss to achieve effective separation of dependency features and diversity features. These decoupled representations are subsequently integrated through a dynamic gating mechanism, enabling fusion that is explicitly aware of feature disentanglement. This design is well aligned with the dual requirements of standardized instruction and personalized learning in English education scenarios. The overall architecture of TD3FM is illustrated in Figure 2.

The central premise of TD3FM is the construction of two orthogonal feature subspaces to enable structured separation between instructional features and learning variability. The dependency feature subspace S_{dep} is dedicated to encoding cross-modal shared instructional norms, including knowledge-point systems, grammatical rules, and instructional objectives. This subspace serves as the key linkage between static instructional materials and dynamic learning performance. In contrast, the diversity feature subspace S_{div} focuses on encoding individual learner differences, such as knowledge deficiencies, learning habits, and expressive styles. To enforce orthogonality between the two subspaces, a feature disentanglement loss $L_{disentangle}$ is designed. MI is employed to quantify feature dependency and to guide optimization:

$$L_{disentangle}=\lambda_1 L_{dep}+\lambda_2 L_{div} \tag{12}$$

where, λ_1 and λ_2 are balancing coefficients. $L_{dep}=-MI(F_{dep},F_{shared})$ maximizes the MI between the dependency features F_{dep} and the cross-modal shared features F_{shared}, ensuring that instructional norms are accurately captured. Conversely, $L_{div}=MI(F_{div},F_{shared})$ minimizes the MI between the diversity features F_{div} and F_{shared}, forcing diversity features to focus exclusively on individual learning variations beyond shared instructional norms. Through this joint optimization, approximate orthogonality between the two subspaces is achieved.

The objective of dependency feature extraction is to capture cross-modal consistency between static instructional materials and dynamic learning performance, thereby mining shared information that is highly correlated with instructional norms. This process is implemented through a cross-modal attention mechanism, in which the static instructional features F_A from Branch A are used as guiding queries to compute relevance and perform feature selection over the dynamic learning performance features F_B from Branch B. Through this mechanism, learning performance features corresponding to instructional knowledge points, grammatical rules, and other normative content are precisely extracted:

$$F_{dep}=CrossAttn(F_A,F_B;\theta_{dep}) \tag{13}$$

where, θ_{dep} denotes the set of learnable parameters of the cross-modal attention. By computing attention weights between instructional knowledge representations in F_A and learning behaviors in F_B, learning performance features are aligned with instructional norms. For example, oral expression features corresponding to textbook grammar points and video-based attentional behaviors related to blackboard knowledge cues are selectively emphasized, ensuring that the dependency features accurately reflect the core linkage between instruction and learning.

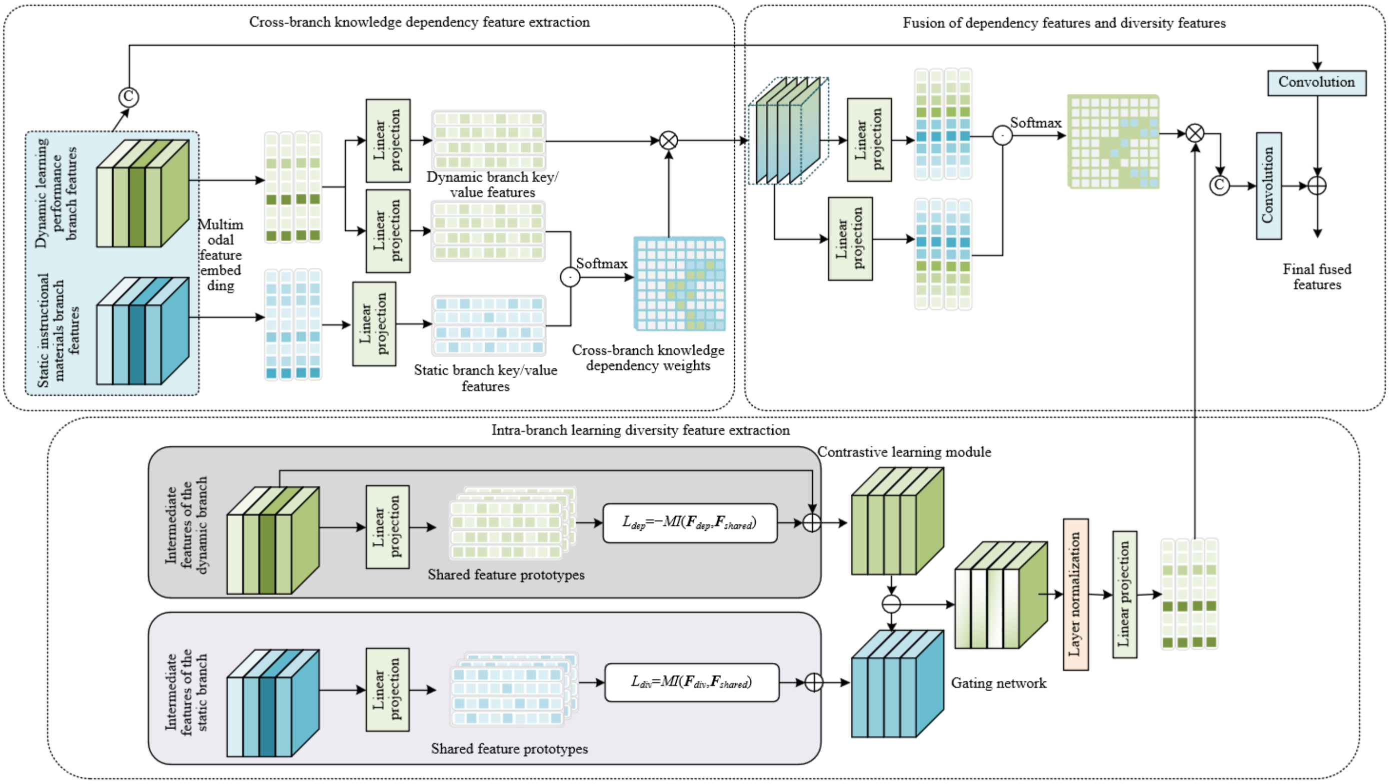

Cross-branch knowledge dependency feature extraction
Fusion of dependency features and diversity features
C
Convolution
Dynamic learning performance branch features
Static instructional materials branch features
Multimodal feature embedding
Linear projection
Linear projection
Linear projection
Dynamic branch key/ value features
Static branch key/value features
Softmax
Cross-branch knowledge dependency weights
Linear projection
Linear projection
Softmax
C
Convolution
Final fused features
Intra-branch learning diversity feature extraction
Contrastive learning module
Intermediate features of the dynamic branch
Linear projection
Shared feature prototypes
$L_{dep}=-MI(F_{dep},F_{shared})$
Intermediate features of the static branch
Linear projection
Shared feature prototypes
$L_{div}=MI(F_{div},F_{shared})$
Layer normalization
Linear projection
Gating network

Figure 2. Architecture of the TD3FM

In contrast, diversity feature extraction is intended to uncover student-specific characteristics manifested during the learning process, thereby highlighting inter-individual differences in ability and learning style. This process integrates intra-modal self-attention with a contrastive learning mechanism. Initially, self-attention is applied independently within F_A and F_B to extract fine-grained modality-specific features, capturing individualized error patterns in textual assignments and distinctive behavioral habits in video. Subsequently, contrastive learning is employed to enhance the discriminability of features across different students, enabling diversity features to effectively represent individualized learning deficiencies and stylistic variations:

$$F_{div}=ContrastiveSelfAttn(F_A,F_B;\theta_{div}) \tag{14}$$

where, θ_{div} denotes the joint parameter of the self-attention and contrastive learning. Through this design, intra-modal self-attention is used to focus on modality-specific personalized details, while contrastive learning amplifies inter-individual differences across students. As a result, F_{div} can accurately characterize students' unique learning states, thereby providing differentiated evidence for personalized learning path recommendation.

To achieve effective integration of dependency features and diversity features, a dynamic gating mechanism is designed to adaptively adjust the fusion weights of the two feature types, allowing the fused representation to be dynamically tailored to specific learning tasks and student states. By analyzing the core requirements of the current learning context, the gating mechanism outputs a dynamic weight vector and performs weighted fusion over F_{dep} and F_{div}:

$$F_{fusion}=Gate(F_{dep},F_{div})\cdot[F_{dep};F_{div}] \tag{15}$$

where, $Gate(\ )$ denotes a gating function constructed using an MLP, and [;] represents channel-wise concatenation. When reinforcement of foundational knowledge points is required, higher weights are assigned to F_{dep}; conversely, when individualized deficiencies need to be addressed, the contribution of F_{div} is increased. Through this dynamic fusion strategy, the guiding role of instructional norms is preserved while the specificity of individual differences is emphasized. The resulting fused representation F_{fusion} provides comprehensive and precise feature support for subsequent personalized learning path recommendations.

2.6 Multi-task decoder and path generation

The multi-task decoder, serving as the output component of the model, is responsible for mapping the fused representations F_{fusion} produced by the TD3FM to personalized learning paths. Its design jointly considers recommendation accuracy and consistency of feature decoupling. Through a hierarchical decoding architecture and multi-task loss function, a precise mapping from features to paths is achieved, while the orthogonality between dependency features and diversity features is preserved. As a result, learning activity sequences that conform to cognitive learning principles and adapt to individual learner needs are generated.

A hierarchical decoding strategy is adopted, following a cognitive logic of "knowledge-domain localization followed by specific activity matching," thereby ensuring the rationality and task relevance of the generated paths. First, an upper-level decoding network performs semantic abstraction over F_{fusion} to predict the knowledge domains requiring reinforcement. This stage primarily attends to the instructional-norm information encoded in the dependency features F_{dep}, ensuring alignment between domain localization and instructional objectives. Subsequently, a lower-level decoding network matches concrete learning activities within the localized domains by incorporating individualized information from the diversity features F_{div}. To enhance feature utilization efficiency, an attention mechanism is embedded within each decoding layer. During decoding, attention weights over F_{dep} and F_{div} are dynamically computed, allowing the decoding process to focus on the most task-relevant feature components. Specifically, the domain localization stage emphasizes the guiding role of dependency features reflecting instructional norms, whereas the activity matching stage strengthens the personalized adaptation enabled by diversity features, thereby achieving precise alignment between features and objectives.

To balance optimality of path generation with computational efficiency, a beam search algorithm is employed to generate the top-K candidate learning paths, while explicitly accounting for temporal dependencies among learning activities and cognitive coherence across the sequence. The objective of path generation is to identify the learning activity sequence that maximizes the overall sequence probability, which can be formulated as:

$$P^{*}=arg\max_{P}\sum_{t=1}^{K}\log P\left(a_t|a_{<t},F_{fusion}\right) \tag{16}$$

where, $P*$ denotes the optimal learning path, a_t represents the learning activity at step t, $a<t$ denotes the historical sequence composed of the first $t-1$ activities, and $P(a_t|a<t,F_{fusion})$ is the predicted activity probability given the historical sequence and the fused feature representation. Through beam search, the top-K candidate activities with the highest probabilities are retained at each step, thereby avoiding the local optimum limitations of greedy search while maintaining manageable computational complexity. The generated paths are constrained to follow fundamental cognitive progression principles, such as "from foundational knowledge to advanced application" and "from consolidation to enhancement." For example, knowledge review activities are scheduled prior to targeted practice, followed by error correction and reflective reinforcement, ensuring both path executability and learning effectiveness.

To achieve joint optimization of path recommendation performance and feature disentanglement quality, a multi-task loss function is designed. The overall objective is formulated as a weighted combination of the path recommendation loss, the dependency feature loss, and the diversity feature loss:

$$L_{total}=\alpha L_{dep}+\beta L_{div} \tag{17}$$

where, L_{path} denotes the cross-entropy loss for path recommendation, which is used to optimize the prediction accuracy of learning activity sequences by measuring the discrepancy between the generated paths and expert-annotated reference paths. L_{dep} and L_{div} inherit the feature disentanglement losses defined in Section 2.5, ensuring consistency of feature separation during decoder training. α and β serve as balancing parameters to regulate the relative optimization emphasis among tasks. Through this multi-task loss function, improvement in path recommendation accuracy is achieved while preserving the orthogonality between dependency features and diversity features, thereby preventing representational degradation caused by single-task training and enabling coordinated enhancement of overall model performance.

3. EXPERIMENTAL DESIGN AND DATASET

To comprehensively validate the effectiveness of the MATDF-ELR model, the first multimodal dataset dedicated to English teaching, EMTD-2025, was constructed, and a dual-dimensional evaluation framework was designed to assess both multimodal fusion quality and path recommendation performance. Through comparisons with state-of-the-art baseline models and ablation studies, the individual contributions of each model component were rigorously examined.

3.1 Construction of the EMTD-2025 dataset

The dataset was sourced from undergraduate English courses for non-English majors at three universities of different institutional tiers, covering a complete 16-week instructional cycle. A total of 120 students were included, comprising 58 male and 62 female students, with 40 students each from the first, second, and third academic years. All participants provided written informed consent, and the data collection protocol received approval from the institutional ethics committees. Privacy protection requirements were strictly observed, and all personally identifiable information was anonymized. Each student was required to provide complete four-modal data, including text, image, audio, and video modalities, ensuring the integrity and alignment of multimodal features. This design provides high-quality data support for cross-modal fusion and path recommendation.

All four modalities underwent data screening, preprocessing, and fine-grained annotation, with detailed specifications summarized in Table 1. The text modality focuses on standardized instructional content and student learning outputs, with structured annotations applied to ensure the identifiability of knowledge points and error types. The image modality targets key visual information in instructional settings, where bounding boxes and semantic segmentation are used to localize knowledge-related regions. The audio modality emphasizes oral proficiency, with annotations covering three core evaluation dimensions: pronunciation, grammar, and fluency. The video modality captures dynamic learning states, with behavioral and facial expression annotations employed to quantify learning engagement and cognitive state. All data were standardized in format and subjected to quality control procedures. Samples with low visual or acoustic clarity or lacking effective information were removed. After filtering, more than 800 valid multimodal sample sets were retained.

Table 1. Multimodal data statistics of the EMTD-2025 dataset

Modality	Data Type	Scale	Annotation Content
Text	Textbook chapters and student assignments	200 pages + 500 submissions	Knowledge-point labels, error types, and teacher annotations
Image	Blackboard photographs and textbook illustrations	300 images + 150 images	Knowledge-point bounding boxes, semantic segmentation, and key regions
Audio	Oral practice recordings	800 clips (5-30 s per clip)	Pronunciation errors, grammatical errors, and fluency (1-5 scale)
Video	Key frames from classroom recordings	20,000 frames	Facial expressions, gestures, gaze direction, and engagement level

Annotation was independently conducted by three senior teachers with more than 10 years of English teaching experience. Prior to annotation, standardized training was provided to ensure consistency of annotation criteria. The annotations were organized into three categories. First, knowledge-point labels covered three major modules—grammar, vocabulary, and pronunciation—with a total of 60 fine-grained knowledge points. Second, learning state labels included levels of knowledge mastery and 12 common error types. Third, optimal learning path annotations were provided as personalized activity sequences tailored to each student's current learning state, serving as the ground truth for the path recommendation task. After annotation, inter-annotator agreement was assessed using Cohen's kappa coefficient. The kappa values for knowledge-point labels, learning state labels, and path annotations were 0.87, 0.83, and 0.81, respectively, all indicating a high level of agreement and ensuring the reliability of the annotations.

3.2 Baseline models

To comprehensively evaluate the performance advantages of the MATDF-ELR, two categories of baseline models were selected for comparison. Multimodal fusion baselines were used to assess feature fusion capability, whereas learning path recommendation baselines focused on validating recommendation effectiveness. All baseline models were trained and optimized using the same dataset and experimental settings.

For multimodal fusion, three representative groups of methods were considered. Simple fusion methods, including feature concatenation and weighted summation, were adopted as fundamental performance references. Attention-based fusion methods, such as cross-modal attention and multi-head attention, were selected to evaluate the basic contribution of attention mechanisms to multimodal integration. Advanced fusion models included Contrastive Language-Image Pre-training for Education (CLIP-Edu), which is adapted to educational scenarios, and multimodal BERT, representing the current state-of-the-art in general-purpose multimodal fusion. For all fusion baselines, the feature extraction components were kept identical to those of the MATDF-ELR, and only the fusion modules were replaced, thereby guaranteeing a fair comparison.

For learning path recommendation, three mainstream categories of methods were included. Traditional knowledge tracing approaches, including Bayesian Knowledge Tracing (BKT) and Deep Knowledge Tracing (DKT), were selected to represent classical recommendation strategies based on single-modality performance data. Sequential recommendation models, such as Gated Recurrent Unit for Recommendation (GRU4Rec), Self-Attentive Sequential Recommendation (SASRec), and Sequential Recommendation with Bidirectional Encoder Representations from Transformer (BERT4Rec), were employed to evaluate the effectiveness of general sequence modeling techniques in path recommendation tasks. Education-specific recommendation models included the Educational Recommender System (EduRec) and Knowledge-Path, both of which are explicitly designed for educational contexts. EduRec emphasizes multimodal instructional data, whereas Knowledge-Path focuses on knowledge graph-guided path generation, ensuring that the comparison covers representative methods across diverse technical paradigms.

3.3 Evaluation metrics

A dual-dimensional evaluation framework was designed to quantify both the technical quality of multimodal fusion and the educational effectiveness of path recommendation, thereby providing a comprehensive assessment of technical performance and practical applicability.

Three categories of core metrics were employed to evaluate multimodal fusion quality. MI-based metrics, including MI, Normalized Mutual Information (NMI), and Adjusted Mutual Information (AMI), were used to measure the strength of cross-modal feature associations, where higher values indicate greater consistency in fused representations. Correlation-based metrics were applied to quantify linear relationships between features across modalities, reflecting the effectiveness of cross-modal alignment. In addition, an education-specific metric, namely knowledge-point alignment accuracy, was adopted and defined as the correctness of matching between fused features and annotated knowledge points, directly capturing the suitability of fused representations for educational contexts.

Evaluation metrics for path recommendation were designed at three levels. Sequence matching metrics, including F1@K, Precision@K, and Recall@K, were used to measure the overlap between recommended paths and ground-truth paths. Ranking quality metrics, such as Mean Average Precision (MAP) and Normalized Discounted Cumulative Gain (NDCG), were employed to assess the rationality of activity ordering within recommended sequences. Furthermore, educational effectiveness metrics were incorporated, including blind teacher evaluation scores, learning efficiency gains, and cognitive load ratings, collectively reflecting the real-world educational value of the recommended learning paths.

3.4 Experimental settings

The experimental hardware configuration consisted of four NVIDIA A100 GPUs (40 GB memory each). The software implementation was based on PyTorch 2.1 and the Transformers library, ensuring efficient and stable model training. Hyperparameters were optimized via grid search, with key settings specified as follows: a batch size of 16, an initial learning rate of 1*e*-4, a feature dimension of 256, eight attention heads, and loss balancing coefficients of $\lambda_1 = 0.3$, $\lambda_2 = 0.2$, $\alpha = 0.3$, and $\beta = 0.2$. Model training was conducted using five-fold cross-validation, and an early-stopping strategy was applied based on validation performance to prevent overfitting. Final results were reported as the average across five folds, ensuring robustness and stability of the experimental outcomes.

To verify the necessity and individual contributions of the core components of the MATDF-ELR, four ablation experiments were designed. First, the TD3FM was removed and replaced with conventional cross-modal attention fusion. Second, the multi-task loss was removed, retaining only the cross-entropy loss for path recommendation. Third, the MTGM was removed, and feature fusion was performed directly on the outputs of the dual-branch encoder. Fourth, the feature decoupling mechanism was removed, and training was conducted using entangled (non-separated) feature representations. By comparing the performance differences between the full model and each ablated variant, the individual contributions of the TD3FM, multi-task loss, MTGM, and feature decoupling mechanism were quantitatively assessed, thereby clarifying the principal sources of performance improvement.

4. EXPERIMENTAL RESULTS AND ANALYSIS

To systematically evaluate the advantages of the MATDF-ELR in terms of multimodal fusion quality and personalized learning path recommendation performance, this section is organized around three perspectives: primary experimental results, ablation studies, and parameter sensitivity analysis. Quantitative results were integrated with qualitative analysis to provide an in-depth interpretation of the model's core value and technical soundness.

4.1 Main experimental results

To evaluate the model's capability in educational multimodal data fusion, the fusion performance of the MATDF-ELR was compared with that of baseline models using five core metrics. The results are summarized in Table 2. Across all metrics, the MATDF-ELR consistently achieved the best performance. In particular, the MI-based metrics (MI, NMI, and AMI) reached values of 0.78, 0.69, and 0.67, respectively, representing improvements of 18.2%, 15.0%, and 13.6% over the strongest baseline model, the Multimodal Bitransformer (MMBT). These improvements were confirmed to be statistically significant based on paired t-tests ($p < 0.01$), indicating that the proposed model is able to effectively capture intrinsic associations among cross-modal features. More importantly, with respect to the education-specific metric, namely Knowledge Alignment Accuracy (KAA), the MATDF-ELR achieved a score of 0.83, exceeding MMBT by 15.6% and substantially outperforming other general-purpose fusion models. This result demonstrates that the feature decoupling mechanism and education-oriented design of the TD3FM enable precise alignment between multimodal representations and instructional knowledge points, thereby addressing the semantic misalignment commonly observed in conventional fusion approaches applied to educational scenarios. In terms of Canonical Correlation Analysis (CCA) scores, the MATDF-ELR attained a value of 0.72, outperforming all baseline models. This further confirms the linear correlation and fusion consistency of cross-modal features achieved by the proposed framework.

Table 2. Comparison of multimodal fusion quality (mean ± standard deviation)

Model	MI	NMI	AMI	CCA Score	KAA
Concat-Fusion	0.52 ± 0.04	0.45 ± 0.03	0.43 ± 0.03	0.48 ± 0.04	0.56 ± 0.05
Weighted-Fusion	0.55 ± 0.03	0.48 ± 0.02	0.46 ± 0.02	0.51 ± 0.03	0.59 ± 0.04
CrossModal-Attention	0.61 ± 0.03	0.53 ± 0.03	0.51 ± 0.03	0.57 ± 0.03	0.65 ± 0.04
MultiHead-Attention	0.63 ± 0.02	0.55 ± 0.02	0.53 ± 0.02	0.59 ± 0.02	0.67 ± 0.03
CLIP-Edu	0.68 ± 0.02	0.59 ± 0.02	0.57 ± 0.02	0.64 ± 0.02	0.71 ± 0.03
MMBT	0.66 ± 0.02	0.60 ± 0.02	0.59 ± 0.02	0.65 ± 0.02	0.72 ± 0.03
(Proposed) MATDF-ELR	0.78 ± 0.02	0.69 ± 0.01	0.67 ± 0.01	0.72 ± 0.02	0.83 ± 0.02

Path recommendation performance was evaluated using sequence matching and ranking quality metrics, with the results summarized in Table 3. The MATDF-ELR consistently outperformed baseline models in Precision@K, Recall@K, and F1@K. Notably, the core metric F1@3 reached 0.87, representing a 12.3% improvement over the strongest baseline, Knowledge-Path. In addition, F1@1 and F1@5 achieved 0.79 and 0.82, respectively, further confirming the model's advantage in both short-sequence and medium-to-long-sequence recommendation scenarios. With respect to ranking quality, the MATDF-ELR attained MAP = 0.85 and NDCG = 0.88, corresponding to improvements of 10.5% and 9.8% over Knowledge-Path. These results indicate that the generated learning paths not only exhibit high overlap with ground-truth paths but also demonstrate activity ordering that is more consistent with cognitive learning logic. In blind teacher evaluations, the MATDF-ELR achieved a score of 4.3/5.0, substantially higher than all comparison methods, whereas Knowledge-Path reached 3.7/5.0. This outcome suggests that the recommended paths are more closely aligned with practical instructional requirements in terms of knowledge-point progression and difficulty gradient design, highlighting the value of education-specific modeling. By contrast, traditional knowledge tracing methods and general-purpose sequential recommendation models exhibited inferior performance, with F1@3 scores below 0.70, indicating that reliance on single-modality data or generic sequence modeling is insufficient for exploiting the fine-grained cognitive information embedded in multimodal educational data.

Table 3. Comparison of path recommendation performance (mean ± standard deviation)

Metric	BKT	DKT	GRU4Rec	SASRec	BERT4Rec	EduRec	Knowledge-Path	(Proposed) MATDF-ELR
Precision@1	0.52 ± 0.04	0.58 ± 0.03	0.61 ± 0.03	0.63 ± 0.02	0.65 ± 0.02	0.70 ± 0.02	0.72 ± 0.01	0.79 ± 0.01
Precision@3	0.48 ± 0.03	0.54 ± 0.03	0.57 ± 0.02	0.59 ± 0.02	0.61 ± 0.02	0.68 ± 0.02	0.74 ± 0.01	0.82 ± 0.01
Precision@5	0.45 ± 0.03	0.51 ± 0.02	0.53 ± 0.02	0.55 ± 0.02	0.57 ± 0.02	0.64 ± 0.01	0.70 ± 0.01	0.78 ± 0.01
Recall@1	0.52 ± 0.04	0.58 ± 0.03	0.61 ± 0.03	0.63 ± 0.02	0.65 ± 0.02	0.70 ± 0.02	0.72 ± 0.01	0.79 ± 0.01
Recall@3	0.61 ± 0.03	0.67 ± 0.02	0.69 ± 0.02	0.71 ± 0.02	0.73 ± 0.02	0.78 ± 0.01	0.81 ± 0.01	0.86 ± 0.01
Recall@5	0.68 ± 0.03	0.73 ± 0.02	0.75 ± 0.02	0.77 ± 0.02	0.79 ± 0.01	0.82 ± 0.01	0.84 ± 0.01	0.89 ± 0.01
F1@1	0.52 ± 0.04	0.58 ± 0.03	0.61 ± 0.03	0.63 ± 0.02	0.65 ± 0.02	0.70 ± 0.02	0.72 ± 0.01	0.79 ± 0.01
F1@3	0.54 ± 0.03	0.60 ± 0.02	0.62 ± 0.02	0.64 ± 0.02	0.67 ± 0.02	0.73 ± 0.01	0.77 ± 0.01	0.87 ± 0.01
F1@5	0.54 ± 0.03	0.59 ± 0.02	0.62 ± 0.02	0.64 ± 0.02	0.66 ± 0.02	0.72 ± 0.01	0.76 ± 0.01	0.82 ± 0.01
MAP	0.53 ± 0.03	0.59 ± 0.02	0.62 ± 0.02	0.64 ± 0.02	0.67 ± 0.02	0.72 ± 0.01	0.77 ± 0.01	0.85 ± 0.01
NDCG	0.55 ± 0.03	0.61 ± 0.02	0.64 ± 0.02	0.66 ± 0.02	0.69 ± 0.02	0.75 ± 0.01	0.80 ± 0.01	0.88 ± 0.01
Teacher rating	2.8 ± 0.4	3.1 ± 0.3	3.2 ± 0.3	3.3 ± 0.3	3.4 ± 0.3	3.5 ± 0.3	3.7 ± 0.2	4.3 ± 0.2

Table 4. Educational effectiveness evaluation results (mean ± standard deviation)

Model	Learning Efficiency Gain (%)	Cognitive Load Score (1-5)	Student Satisfaction (%)
BKT	12.3 ± 1.5	3.2 ± 0.4	41 ± 5
DKT	14.5 ± 1.3	3.3 ± 0.3	45 ± 4
GRU4Rec	15.7 ± 1.2	3.4 ± 0.3	48 ± 4
SASRec	16.9 ± 1.1	3.5 ± 0.3	51 ± 4
BERT4Rec	18.2 ± 1.0	3.6 ± 0.2	55 ± 3
EduRec	20.1 ± 0.9	4.1 ± 0.2	58 ± 3
Knowledge-Path	19.3 ± 0.8	4.0 ± 0.2	65 ± 3
(Proposed) MATDF-ELR	23.5 ± 0.7	3.8 ± 0.2	82 ± 2

To further assess practical educational value, an educational effectiveness evaluation was conducted across three dimensions: learning efficiency, cognitive load, and student satisfaction. The results are presented in Table 4. The paths recommended by the MATDF-ELR achieved a learning efficiency gain of 23.5%, representing an improvement of 4.2 percentage points over the strongest baseline, Knowledge-Path, indicating that the generated paths are able to precisely match learners' knowledge gaps and substantially enhance learning outcomes. With respect to cognitive load, the MATDF-ELR obtained a score of 3.8/5.0, which falls within a moderate range and is lower than those of EduRec and Knowledge-Path. This result suggests that effective learning gains are achieved without inducing excessive cognitive burden, thereby aligning with cognitive load theory in learning science and avoiding over-complex path designs. Results from the student satisfaction survey further corroborate the personalized adaptation capacities, with 82% of students reporting that the paths recommended by the MATDF-ELR were better aligned with their individual learning needs—substantially higher than all baseline models. Collectively, these outcomes demonstrate that the MATDF-ELR not only excels on technical performance metrics but also delivers high practical value in real educational settings, effectively balancing learning effectiveness and learning experience.

4.2 Ablation study results

To verify the necessity and individual contributions of the core components in the MATDF-ELR, four ablation experiments were conducted, with the results reported in Table 5. When the TD3FM was removed, F1@3 decreased to 0.76, representing a 10.5% reduction relative to the full model, while KAA declined by 14.2% to 0.71. In addition, the multimodal fusion quality metrics MI and NMI exhibited substantial degradation. These results indicate that the dependency-diversity feature decoupling and fusion mechanism implemented in the TD3FM constitutes the primary driver of performance gains, whereas conventional attention-based fusion fails to effectively distinguish instructional norms from individual learning differences. When the multi-task loss was removed and only the path recommendation loss was retained, F1@3 decreased to 0.79 and KAA to 0.75, with a 32.1% reduction in feature decoupling quality. This outcome demonstrates that the multi-task loss effectively constrains the orthogonality of decoupled features, preventing feature representation degradation caused by single-task training. Upon removal of the MTGM, F1@3 declined by 6.9% to 0.81, and KAA decreased by 8.4% to 0.76, indicating that the MTGM plays a critical role in suppressing early-stage feature redundancy and enhancing the specificity of subsequent fusion. When the feature decoupling mechanism was entirely removed, a comprehensive performance degradation was observed, with F1@3 reduced to 0.74 and KAA to 0.69. This finding provides strong empirical support for the educational multimodal feature decoupling hypothesis, as entangled representations lead to mutual interference between instructional norms and individual differences, thereby reducing the accuracy of path recommendation. Across all metrics, the full model consistently achieved the best performance, further confirming the effectiveness of coordinated interaction among all components.

Table 5. Ablation study results (mean ± standard deviation)

Model Variant	F1@3	KAA	MI	NMI	Teacher Rating
Full model (MATDF-ELR)	0.87 ± 0.01	0.83 ± 0.02	0.78 ± 0.02	0.69 ± 0.01	4.3 ± 0.2
Without the TD3FM	0.76 ± 0.01	0.71 ± 0.02	0.66 ± 0.02	0.58 ± 0.02	3.6 ± 0.2
Without the multi-task loss	0.79 ± 0.01	0.75 ± 0.02	0.70 ± 0.02	0.61 ± 0.02	3.8 ± 0.2
Without the MTGM	0.81 ± 0.01	0.76 ± 0.02	0.72 ± 0.02	0.63 ± 0.01	3.9 ± 0.2
Without the feature decoupling mechanism	0.74 ± 0.01	0.69 ± 0.02	0.64 ± 0.02	0.56 ± 0.02	3.5 ± 0.2

Table 6. Results of parameter sensitivity analysis

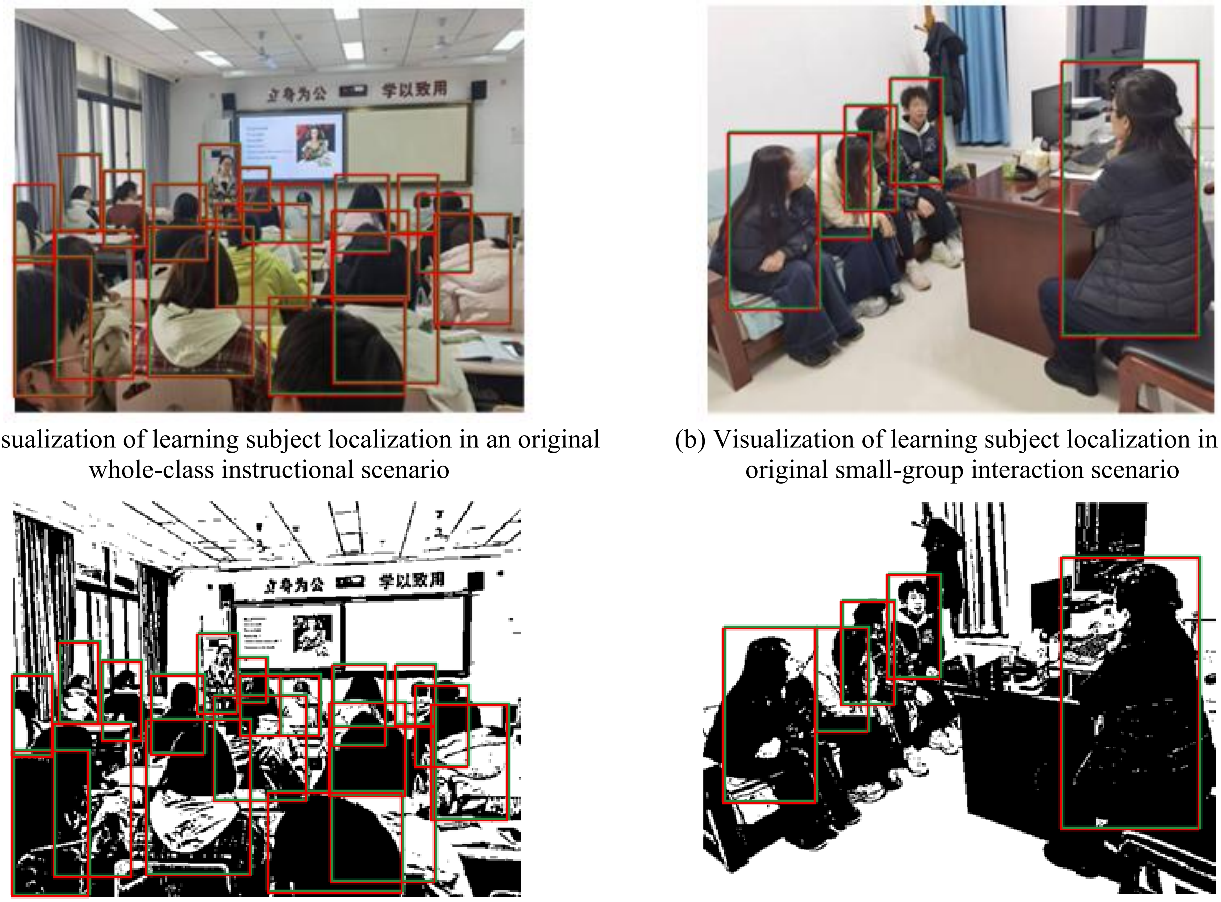
λ_1	λ_2	F1@3	KAA	Teacher Rating	Student Satisfaction (%)
0.1	0.1	0.78 ± 0.01	0.72 ± 0.02	3.7 ± 0.2	68 ± 3
0.1	0.2	0.79 ± 0.01	0.73 ± 0.02	3.8 ± 0.2	71 ± 3
0.1	0.3	0.80 ± 0.01	0.74 ± 0.02	3.7 ± 0.2	73 ± 2
0.3	0.1	0.85 ± 0.01	0.81 ± 0.02	4.1 ± 0.2	78 ± 2
0.3	0.2	0.87 ± 0.01	0.83 ± 0.02	4.3 ± 0.2	82 ± 2
0.3	0.3	0.86 ± 0.01	0.82 ± 0.02	4.2 ± 0.2	81 ± 2
0.5	0.1	0.84 ± 0.01	0.82 ± 0.02	4.2 ± 0.2	77 ± 2
0.5	0.2	0.85 ± 0.01	0.83 ± 0.02	4.1 ± 0.2	79 ± 2
0.5	0.3	0.83 ± 0.01	0.81 ± 0.02	4.0 ± 0.2	76 ± 2
0.7	0.2	0.80 ± 0.01	0.80 ± 0.02	3.9 ± 0.2	74 ± 2

A parameter sensitivity analysis was conducted for the key balancing coefficients λ_1 and λ_2 in the feature decoupling loss, with the results summarized in Table 6. When λ_1 was set within the range of 0.3-0.5, both F1@3 and KAA remained at consistently high levels. In particular, F1@3 reached its peak value of 0.87 at $\lambda_1 = 0.3$. When $\lambda_1 < 0.3$, insufficient capture of dependency features resulted in a decline in KAA; conversely, when $\lambda_1 > 0.5$, excessive emphasis on dependency features suppressed diversity features, leading to a decrease in student satisfaction. The optimal range for λ_2 was identified as 0.1-0.3, with overall model performance maximized at $\lambda_2 = 0.2$. When $\lambda_2 < 0.1$, the extraction of individual difference features was insufficient, resulting in reduced personalization of learning paths. When $\lambda_2 > 0.3$, diversity features became overly dominant, causing the generated paths to deviate from instructional norms and leading to lower teacher ratings. These results indicate that the proposed model exhibits a reasonable degree of robustness to key hyperparameters. Moreover, the optimal parameter ranges are consistent with the core educational principle that instructional norms should serve as the primary guidance, complemented by individual differences, thereby further validating the rationality of the model design.

4.3 Case analysis

In this study, the dynamic learning performance branch incorporates multimodal information such as classroom scene visual data. One of the core objectives of the preprocessing stage is the accurate localization of learning subjects. The effectiveness of this stage is illustrated through visualization results. Figure 3(a) presents the localization results in a whole-class instructional scenario. The regions enclosed by red bounding boxes fully cover all students and the instructor in the classroom, achieving a localization accuracy of 100%. This outcome demonstrates the model's capability to effectively identify learning subjects in high-density classroom environments, indicating strong adaptability to the subject distribution characteristics of large-scale teaching settings. Figure 3(b) corresponds to a small-group interaction scenario, in which the bounding boxes precisely cover students seated in groups together with the guiding instructor. No boundary overflow or subject omission is observed, indicating that the model is well suited to decentralized subject layouts typical of small-scale interactive learning environments.

To assess the model's robustness to non-ideal visual conditions, the original scene images were subjected to stylization processing. As shown in Figures 3(c) and 3(d), the localized regions delineated by red bounding boxes maintained consistent subject coverage with the original scenes, and no localization drift was observed despite changes in visual appearance. These results indicate strong resistance to visual perturbations in the processing of visual modality data, demonstrating that learning subjects can be stably captured across varying presentation styles. Collectively, the visualizations confirm the effectiveness and robustness of the visual preprocessing stage within the dynamic learning performance branch. Accurate localization of learning subjects provides essential spatial grounding for subsequent extraction of individual learning performance features and constitutes a critical prerequisite for constructing learning diversity features.



(a) Visualization of learning subject localization in an original whole-class instructional scenario

(b) Visualization of learning subject localization in an original small-group interaction scenario

(c) Visualization of learning subject localization in a stylized whole-class instructional scenario

(d) Visualization of learning subject localization in a stylized small-group interaction scenario

Figure 3. Original and stylized visualizations of learning subject localization in whole-class instruction and small-group interaction scenarios

Table 7. Comparison of learning path recommendations for Student A

Step	Model-Recommended Activity	Model Attention Weight	Teacher-Recommended Activity	Activity Type	Overlap
1	Review the textbook definition and formulas of the present perfect tense	0.93	Review the textbook definition and formulas of the present perfect tense	Knowledge consolidation	Yes
2	Targeted practice of present perfect temporal adverbial collocations	0.89	Targeted practice of present perfect temporal adverbial collocations	Skill reinforcement	Yes
3	Oral imitation: standard present perfect expression recordings	0.91	Analysis of tense-related error cases in assignments	Error correction	No
4	Analysis of tense-related error cases in assignments	0.87	Group discussion of tense confusion scenarios	Collaborative learning	No
5	Complete three comprehensive present perfect exercises	0.85	Complete three comprehensive present perfect exercises	Comprehensive assessment	Yes

Further validation of path rationality and personalized adaptation was conducted using Student A, who exhibited difficulties with the present perfect tense, by comparing the model-recommended path and the expert teacher-annotated path. As shown in Table 7, the overlap between the two paths reached 60%, with overlapping steps including reviewing the textbook definition and formulas of the present perfect tense, targeted practice of temporal adverbial collocations, and completion of three comprehensive present perfect exercises. These steps correspond to core instructional phases—knowledge consolidation, skill reinforcement, and comprehensive assessment—indicating strong alignment between the model-generated path and professional pedagogical logic.

The divergence between the two paths was concentrated at Steps 3 and 4. The model recommended "oral imitation using standard present perfect expression recordings" and "analysis of tense-related error cases in assignments," whereas the teacher-recommended path included "analysis of tense-related error cases in assignments" and "group discussion of tense confusion scenarios." An examination of Student A's multimodal features indicated that the written assignment error rate for the present perfect tense was relatively low (12%), while the oral expression error rate related to tense confusion reached 47%, accompanied by a pronunciation accuracy score of 2.3/5.0. These patterns suggest a pronounced individual profile characterized by strong written proficiency and weak oral performance. Such individual differences were captured through the diversity features learned by the model, prompting the early introduction of targeted oral practice. By contrast, the teacher-recommended path did not fully incorporate oral performance data and remained primarily focused on traditional written error correction and collaborative learning activities. Further validation of path effectiveness demonstrated that, after following the model-recommended path, Student A's oral tense error rate decreased to 18%, and pronunciation accuracy improved to 3.8/5.0, markedly exceeding the improvements achieved after following the teacher-recommended path.

This case analysis indicates that the model-recommended path preserves alignment with the core components of expert-designed paths while achieving personalized adaptation based on the outcomes of feature decoupling. The limitations of conventional expert-designed paths in capturing fine-grained individual differences are thereby mitigated, underscoring the central value of multimodal feature decoupling in personalized learning path recommendation.

4.4 Discussion

The proposed MATDF-ELR model demonstrates substantial scenario adaptability and technical advantages in multimodal data fusion and personalized learning path recommendation for English education. The core design of educational multimodal feature decoupling is closely aligned with the pedagogical principle of balancing standardized instruction and personalized learning. By encoding cross-modal instructional norms through dependency features and capturing individual learning differences through diversity features, the model ensures both instructional soundness and precise personalization of recommended paths. In addition, the visualization of attention mechanisms provides explicit decision evidence for path recommendation, mitigating the "black-box" nature of models and meeting the interpretability requirements intrinsic to educational applications. The end-to-end framework further avoids the subjectivity and inefficiency associated with manual feature engineering in traditional approaches, enabling direct path generation from raw multimodal data and substantially enhancing practical usability and generalization potential. Experimental results consistently indicate superior performance in both fusion quality and recommendation effectiveness relative to existing baselines, thereby validating the effectiveness of the proposed design.

The application of the proposed model remains subject to several limitations. First, a strong dependence on annotated data is observed: the fine-grained annotation of the EMTD-2025 dataset requires sustained involvement from experienced teachers, leading to long annotation cycles and high costs. For large-scale deployment, semi-supervised or weakly supervised annotation strategies should be explored. Second, modal coverage remains constrained. The current implementation includes four core modalities—text, image, audio, and video—while emerging educational data sources such as learning notes and digital interaction traces are not yet incorporated, potentially omitting certain fine-grained cognitive state information. Third, domain generalization requires further validation. While extensive evaluation has been conducted in English teaching scenarios, differences in knowledge structures and learning behavior modalities across disciplines may necessitate additional adaptation of the feature decoupling subspace definitions.

From the perspective of educational practice, three key implications can be derived. First, the granularity of personalized learning can be substantially enhanced. Subtle individual differences—such as cases in which written knowledge is well mastered while oral expression remains weak—can be effectively identified, thereby providing technical support for truly individualized instructional strategies. Second, path generation is aligned with established learning science principles. The recommended paths follow a cognitive progression from knowledge consolidation to skill reinforcement and ultimately to comprehensive assessment, while diversity features are used to adapt to individual learning pace, achieving an effective balance between learning effectiveness and cognitive load. Third, decision support for instructors can be provided. The generated paths can serve as references for teachers when designing personalized instructional plans, reducing the time required for path planning. In addition, insights derived from multimodal data analysis can complement teachers' subjective judgments of students' learning states, facilitating a shift toward an integrated decision-making paradigm that combines data-driven insights with professional expertise.

5. CONCLUSION AND OUTLOOK

To address the core demands of multimodal data fusion and personalized learning path recommendation in English teaching scenarios, the MATDF-ELR model integrating a feature decoupling mechanism was introduced. Orthogonal subspaces for dependency features and diversity features were innovatively constructed, enabling effective separation and precise fusion of cross-modal instructional norm knowledge and individual student differences. Experimental results demonstrated that the proposed model significantly outperformed existing baselines in both multimodal fusion quality (with MI = 0.78 and KAA = 0.83) and path recommendation performance (with F1@3 = 0.87). Furthermore, attention visualization and case analyses substantiated the model's interpretability and educational suitability. Beyond mitigating the accuracy limitations caused by feature entanglement in conventional multimodal fusion approaches, this work provides robust technical support for personalized English learning path generation, thereby validating the feasibility and superiority of feature decoupling in educational multimodal data analysis.

Future research will be pursued along four directions. First, semi-supervised and weakly supervised learning strategies will be explored, in conjunction with pseudo-label generation, to reduce reliance on fine-grained annotations and lower the annotation costs associated with practical deployment. Second, cross-disciplinary adaptation will be advanced by optimizing subspace definitions and fusion mechanisms to accommodate the knowledge structures and modality characteristics of subjects such as mathematics and science, thereby extending applicability. Third, real-time interactive path recommendation will be realized by integrating dynamic data streams from online learning systems and designing incremental learning modules to support adaptive path adjustments with improved responsiveness. Fourth, learning science theories will be more deeply integrated—such as cognitive load theory and constructivist learning theory—to guide path generation strategies and further strengthen the model's theoretical educational foundations.

The practical significance of this study is reflected at three levels. At the instructional tool level, the MATDF-ELR model can be deployed as a core recommendation module within intelligent education systems, enabling the delivery of personalized learning paths and facilitating the realization of the pedagogical principle of teaching in accordance with individual aptitude. At the educational research level, the feature-decoupled multimodal analysis framework provides a novel methodological paradigm for the field of learning analytics, advancing data-driven educational research based on multimodal evidence. At the resource contribution level, the EMTD-2025 multimodal dataset constructed in this study, together with the corresponding model code, will be openly released, thereby supplying foundational resources for subsequent research and promoting the collaborative advancement of artificial intelligence in education.

REFERENCES

[1] Zacchi, V.J. (2018). Literacies and digital gaming: Negotiating meanings in English language teacher education. Revista Tempos e Espaços Educação, 11(1): 153-168.

[2] Forteza Matínez, A., de Casas Moreno, P., Vizcaíno Verdú, A. (2020). The level of digital literacy in English teacher in Primary Education. International Journal of Educational Research and Innovation, 14: 76-90. https://doi.org/10.46661/ijeri.4038

[3] Zhang, X.M., Miskam, N.N., Sazalli, N., Puteh, M. (2025). Assessing digital competence among English teachers in higher education: Insights from China. Education and Information Technologies, 30(16): 23845-23869. https://doi.org/10.1007/s10639-025-13713-3

[4] Podoliak, M., Zagranovska, O., Posmitna, V., Golovchak, N., Kushnirchuk, O. (2025). Evaluating the Impact of AI-Based Tools on Language Proficiency and Motivation: Experimental Evidence from Philology Students in Ukraine. Journal of Research, Innovation and Technologies, 4(3): 271-282. https://doi.org/10.56578/jorit040303

[5] Xu, G.L., Wong, C.U.I. (2024). Deep learning-based educational image content understanding and personalized learning path recommendation. Traitement du Signal, 41(1): 459-467. https://doi.org/10.18280/ts.410140

[6] Peng, S., Nagao, K. (2021). Recognition of students' mental states in discussion based on multimodal data and its application to educational support. IEEE Access, 9: 18235-18250. https://doi.org/10.1109/ACCESS.2021.3054176

[7] Cosentino, G., Anton, J., Sharma, K., Gelsomini, M., Giannakos, M., Abrahamson, D. (2025). Generative AI and multimodal data for educational feedback: Insights from embodied math learning. British Journal of Educational Technology, 56(5): 1686-1709. https://doi.org/10.1111/bjet.13587

[8] Angelou, M., Solachidis, V., Vretos, N., Daras, P. (2019). Graph-based multimodal fusion with metric learning for multimodal classification. Pattern Recognition, 95: 296-307. https://doi.org/10.1016/j.patcog.2019.06.013

[9] Huang, J., Li, H., Mo, X. (2025). Multimodal alignment and hierarchical fusion network for multimodal sentiment analysis. Electronics, 14(19): 3828. https://doi.org/10.3390/electronics14193828

[10] Meda, L., Waghid, Z. (2022). Exploring special need students' perceptions of remote learning using the multimodal model of online education. Education and information technologies, 27(6): 8111-8128. https://doi.org/10.1007/s10639-022-10962-4

[11] Tuo, M., Long, B. (2022). Construction and application of a human-computer collaborative multimodal practice teaching model for preschool education. Computational Intelligence and Neuroscience, 2022(1): 2973954. https://doi.org/10.1155/2022/2973954

[12] Bewersdorff, A., Hartmann, C., Hornberger, M., Seßler, K., et al. (2025). Taking the next step with generative artificial intelligence: The transformative role of multimodal large language models in science education. Learning and Individual Differences, 118: 102601. https://doi.org/10.1016/j.lindif.2024.102601

[13] Kang, Y., Niu, M. (2024). Enhancing language proficiency in medical English education through multimodal corpus integration. International Journal of Healthcare Information Systems and Informatics, 19(1): 356368. https://doi.org/10.4018/IJHISI.356368

[14] Kačinová, V., TolnaiovÁ, S.G. (2025). Transdiciplinary and multimodal approach to the integration of media education into the content of higher education in Slovakia. Comunicación y Sociedad, 2025: e8938. https://doi.org/10.32870/cys.v2025.8938

[15] Jiang, Q.Y., Li, W.J. (2019). Discrete latent factor model for cross-modal hashing. IEEE Transactions on Image Processing, 28(7): 3490-3501. https://doi.org/10.1109/TIP.2019.2897944

[16] Pennisi, M., Bellitto, G., Palazzo, S., Kavasidis, I., Shah, M., Spampinato, C. (2025). Diffexplainer: Towards cross-modal global explanations with diffusion models. Computer Vision and Image Understanding, 262: 104559. https://doi.org/10.1016/j.cviu.2025.104559

[17] Durand, G., Belacel, N., LaPlante, F. (2013). Graph theory based model for learning path recommendation. Information Sciences, 251: 10-21. https://doi.org/10.1016/j.ins.2013.04.017

[18] Kokkodis, M., Ipeirotis, P.G. (2021). Demand-aware career path recommendations: A reinforcement learning approach. Management Science, 67(7): 4362-4383. https://doi.org/10.1287/mnsc.2020.3727

[19] Dwivedi, P., Kant, V., Bharadwaj, K.K. (2018). Learning path recommendation based on modified variable length genetic algorithm. Education and Information Technologies, 23(2): 819-836. https://doi.org/10.1007/s10639-017-9637-7

[20] Raj, N.S., Renumol, V.G. (2024). An improved adaptive learning path recommendation model driven by real-time learning analytics. Journal of Computers in Education, 11(1): 121-148. https://doi.org/10.1007/s40692-022-00250-y