Cross-Modal Retrieval And Personalized Recommendation For Integrated Innovation, Entrepreneurship, And Moral Education

Zijin Li and Weijie Zhao*

Pingdingshan University, Pingdingshan Henan 467000, China

** Corresponding author. E-mail: keyanzwj@163.com*

Received: May. 04, 2026; Accepted: July. 20, 2026

In the context of the knowledge economy and the transformation of higher education, cultivating innovation, entrepreneurship, and moral responsibility has become a strategic priority. Existing innovation and entrepreneurship education platforms often fail to address diverse student profiles and rarely integrate moral education into personalized learning pathways. This study proposes an integrated teaching platform that unifies innovation, entrepreneurship, and moral education through multimedia networks and neural network-driven cross-modal semantic retrieval. The platform consists of a student learning space, teacher management modules, a multimedia resource repository, and real-time feedback mechanisms. A hybrid neural network model is introduced to map multimodal educational resources and student submissions into a shared semantic space using weak semantic label generation, bidirectional feature crossing, attention-enhanced GRU modules, and position encoding. In this framework, weak semantic labels are automatically generated pseudo-labels derived from semantic category probability distributions of unlabeled multimodal samples and are iteratively incorporated into model training to enhance semantic representation learning. This framework enables personalized content recommendation, dynamic interest tracking, and moral dilemma feedback. Experiments conducted on the Wikipedia, Wikipedia-CNN, NUS-WIDE, and domain-specific I&E-EDU datasets demonstrate that the proposed method consistently outperforms baseline approaches, including MRCR-SNN. On the I&E-EDU dataset, the proposed model improves mean Average Precision (mAP) from 28.3% to 32.7% (+4.4 percentage points), Precision@10 from 43.1% to 49.4% (+6.3 percentage points), and increases student engagement by 26.5%.

Keywords: Cross-modal retrieval; Personalized recommendation; Entrepreneurship education; Moral education; Multimodal neural networks.

http://dx.doi.org/10.6180/jase.202611_34.048

1. Introduction

In the era of digital transformation and knowledge economy, higher education emphasizes innovation, entrepreneurship, and moral development beyond theory, fostering creativity and problem-solving [1]. Integrated innovation and entrepreneurship education (IEE) aligns with national talent strategies. Advances in multimedia networks and AI/deep learning enable personalized, adaptive, interactive teaching platforms [2, 3]. Despite progress in IEE, current teaching platforms still struggle to support individual student needs and engagement. A key limitation is the lack of scalable intelligent systems aligning multimodal resources with diverse learning behaviors and moral development trajectories [4]. Traditional platforms use static delivery, ignoring dynamic student profiles, leading to low engagement. Moral education integration remains fragmented [5, 6].

Recent studies have applied advanced technologies in education, including organizational innovation [7] and bib-

liometric analysis of the digital economy [8]. Multimedia-based teaching models [9, 10] improved engagement. CNN, RNN, and attention mechanisms enable multimodal processing and cross-modal retrieval. These advances support AI recommendation systems such as Deep&Cross [11], GRU4Rec [12], and LightSANs [13]. Collectively, research in educational technology has progressively evolved from conventional e-learning systems toward intelligent recommendation, multimodal learning, and cross-modal retrieval frameworks. Building upon these developments, the proposed methodology integrates semantic retrieval, weak labeling, and personalized recommendation to address emerging educational requirements.

Nevertheless, existing approaches have notable limitations in integrated innovation and entrepreneurship education (IEE) with moral education. They often model student behaviors and learning resources separately, ignoring semantic alignment and temporal evolution of interests [14, 15]. Weakly labeled or unlabeled data are underutilized, limiting learning efficiency. Many methods rely on shallow fusion, failing to capture complex cross-modal relationships needed for personalization. Moreover, prior studies are typically evaluated on generic e-learning systems, lacking alignment with IEE-specific pedagogical and moral objectives. To address these gaps, this paper proposes an integrated teaching platform for higher education that combines innovation, entrepreneurship, and moral education within a multimedia network framework. The core of the proposed platform is a neural network-based cross-modal retrieval model, designed to project multimodal educational materials and student submissions into a shared semantic space where their relevance and alignment can be effectively computed. Specifically, the contributions of this research are fourfold:

(1) We design a modular teaching platform integrating student, teacher, resource, and management spaces, enabling real-time data collection, dynamic interest modeling, and moral reasoning support.

(2) We propose a weak semantic label generation mechanism using unlabeled multimodal student data to improve representation learning in sparse labeling settings.

(3) We develop a hybrid neural model combining position encoding, bidirectional feature crossing, AGRU attention, and cross-modal projection for aligned text-image semantic representation and personalized recommendation.

(4) We evaluate the proposed system on benchmark datasets (Wikipedia, NUS-WIDE, Wikipedia-CNN) and a domain-specific I&E-EDU dataset, demonstrating improved mAP, recall, and student engagement over baselines such as MRCR-SNN, GRU4Rec, and Deep&Cross.

2. Materials and methods

System architecture and design

Overall architecture of the platform

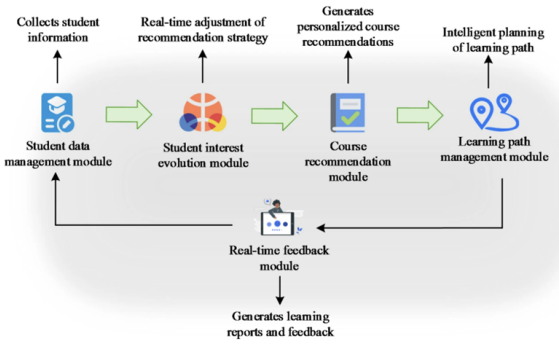

Fig. 1. Overall architecture of the integrated teaching platform

The integrated teaching platform is a modular intelligent system using multimedia networks and recommendation technologies. Fig. 1 shows five interacting modules. The Student Data Management Module maintains student profiles, formally defined as:

$$x_i = [x_{i1}, x_{i2}, \ldots, x_{id}] \tag{1}$$

Where d represents the number of feature dimensions and ii indexes the student. Based on the collected student profile x_i, the Learning Path Management Module intelligently plans a personalized learning path, denoted by the sequence:

$$P_i = \{u_1, u_2, \ldots, u_T\} \tag{2}$$

where T is the total number of learning units planned for student i. This planned path aims to optimize alignment with the student's interests, moral education objectives, and innovationentrepreneurship competency development. Entrepreneurial competencies, including innovation capability, opportunity recognition, problem-solving, leadership, and decisionmaking skills, are prioritized during personalized learning sequence construction. These competency indicators guide adaptive pathways, strengthening entrepreneurship development while aligning with innovation-oriented educational objectives.

The Course Recommendation Module optimizes personalized recommendations, selecting the most suitable course c_i^* for each student using profile x_i and path P_i :

$$c_i^* = \text{argmax}_{c \in C} U(c, x_i, P_i) \tag{3}$$

where c_i^* denotes the optimal recommended course for student i, C denotes the set of available candidate courses, x_i represents the feature vector of student i, P_i denotes the personalized learning path of student i, and $U(c, x_i, P_i)$ is the predicted utility score function incorporating relevance, expected learning gain, and moral value. The utility function evaluates semantic relevance, expected learning gain, and moral alignment, selecting the course with the highest utility score.

The Student Interest Evolution Module dynamically updates the student's latent interest state vector at each time step t as interactions and motivations evolve, modeled as:

$$z_i(t+1) = f(z_i(t), \Delta L_t) \tag{4}$$

where ΔL_t represents the learning gain at time t , and $f(\cdot)$ is a non-linear update function designed to reflect evolving preferences. The transformation function adaptively updates the student interest state by combining historical preferences with new learning gains, enabling personalized modeling that captures evolving engagement patterns, learning progress, and behavioral changes in dynamic environments.

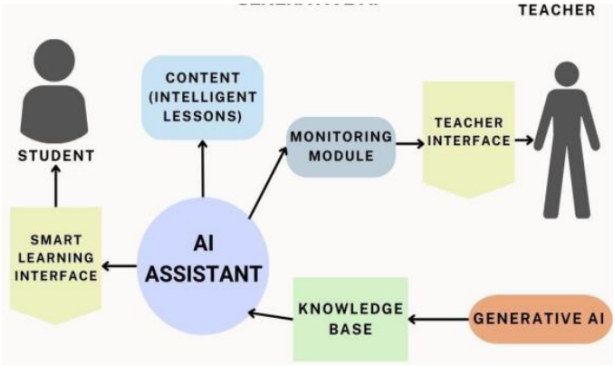

Fig. 2. Educational system architecture

Fig. 2 illustrates four components-repository, teaching management, student space, and teacher space-enabling personalized learning and feedback.

Fig. 3 shows interactions among student, smart interface, AI assistant, knowledge base, and generative AI, enabling continuous Q&A, lesson delivery, and result display until completion.

Workflow of teaching and learning

Fig. 4 illustrates the teaching-learning workflow. The DeepSeek module extracts microlecture topics T_w (written text) and video scripts T_s (spoken text):

$$T_w, T_s = \text{DeepSeek}(M) \tag{5}$$

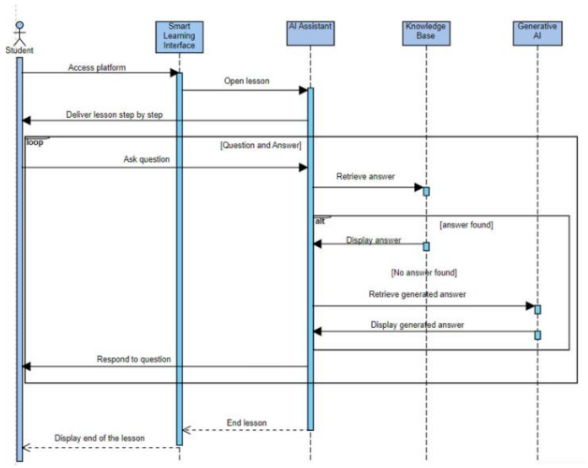

Fig. 3. Sequence diagram of the student-platform interaction

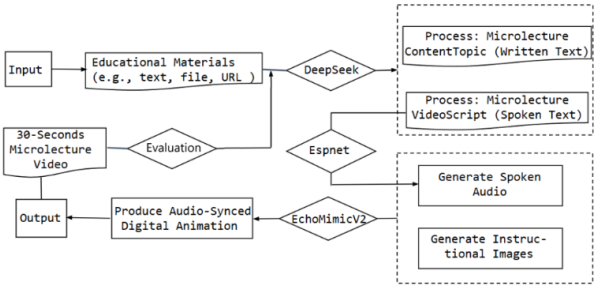

Fig. 4. Workflow of teaching and learning in the integrated platform

DeepSeek-generated topic and script representations are projected into a shared semantic space, enabling multimodal response encoding, semantic matching, relevance assessment, and feedback generation. Using contextual language modeling, DeepSeek extracts key concepts, semantic relationships, and instructional intentions, producing richer semantic representations, reducing ambiguity, and improving multimodal matching and retrieval accuracy throughout the learning process. Students submit multimodal responses $R = \left\{ r^{(\text{text})}, r^{(\text{audio})}, r^{(\text{image})} \right\}$, matched with T_w and T_s; Espnet and EchoMimicV2 generate synchronized 30-second microlecture videos:

$$V = \text{Espnet}(T_S) + \text{EchoMimicV}\,2(T_w) \tag{6}$$

where V denotes the final video with synchronized audio and digital animation. Generative AI modules continuously enrich the instructional repository by generating text, speech, and animated content from evolving materials, ensuring resource freshness, content diversity, pedagogical adaptability, and continuous enhancement of the integrated

educational ecosystem. This output is delivered to the students, who receive real-time feedback F_i based on their submissions:

$$F_i = f\left(R_i, V\right) \tag{7}$$

where f$(\cdot)$ evaluates student submission R_i against content V, enabling multimodal learning and intelligent assessment. Module interactions follow a closed-loop framework where learning records synchronize student profiles, recommendation outputs are dynamically updated, and user feedback drives iterative learning cycles, continuously refining personalized learning pathways through information exchange.

3. Methodology

Neural network-based cross-modal retrieval model

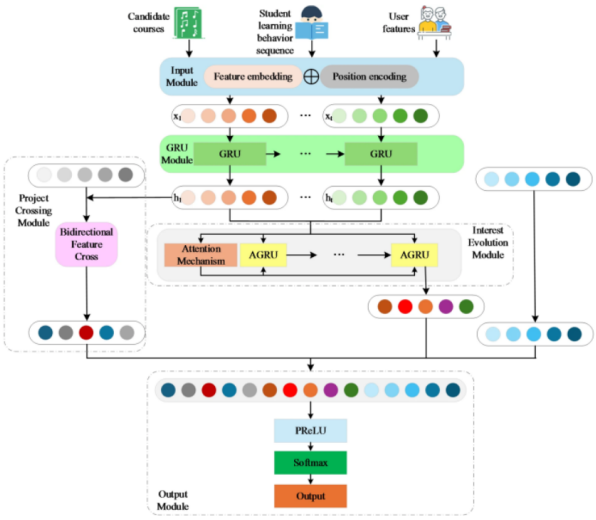

Fig. 5. Architecture of the proposed neural network-based cross-modal retrieval model

Fig. 5 illustrates the proposed cross-modal retrieval model using user features u, lbehavior sequence $\{x_1, x_2, \ldots, x_t\}$, and candidate courses c, embedded by e_i and position encoding p_i :

$$z_i = \text{Embed}\left(x_i\right) + \text{PosEnc}(i) \tag{8}$$

The feature embedding layer maps heterogeneous multimodal inputs into a shared 128-dimensional space, while sinusoidal position encoding preserves temporal order. Together, they provide structured sequential representations for capturing contextual dependencies and evolving learning patterns. A maximum sequence length of 100 interactions supports variable behaviors, stabilizes cross-modal features, and improves sequential representation robustness. The embedded sequence $\{z_1, \ldots, z_t\}$ is then fed into the GRU Module, which captures sequential dependencies and outputs hidden states $\{h_1, \ldots, h_t\}$:

$$h_i = GRU\left(z_i, h_{i-1}\right) \tag{9}$$

The GRU module preserves long-term dependencies by propagating historical hidden states across successive interactions, enabling accumulated learning preferences and gradual shifts in student interests to be captured. Consequently, prolonged platform engagement facilitates stable and representative interest evolution modeling. Hidden states are processed by the Interest Evolution Module using AGRU, where attention weights α_i are computed as:

$$\alpha_i = \frac{\exp\left(q^\top h_i\right)}{\sum_j \cdots \exp\left(q^\top h_j\right)} \tag{10}$$

where α_t denotes the attention weight associated with hidden state h_i, h_i represents the hidden state generated by the GRU layer, q is a learnable query vector, and the denominator normalizes the attention scores across all hidden states.

In the AGRU framework, attention weights dynamically regulate hidden-state updates by emphasizing semantically important learning behaviors and suppressing less relevant interactions. This enables evolving interest representations to capture salient sequential patterns and changes in student preferences over time. Attention coefficients act as adaptive weighting factors in hidden-state updates, assigning greater influence to informative behaviors. Higher coefficients amplify important temporal features, while lower coefficients suppress less informative interactions, improving discriminative sequential representations and learner interest evolution modeling. Concurrently, candidate course features c and user features u undergo Bidirectional Feature Cross, producing a set of crossed features f_{cross} . The bidirectional feature crossing mechanism models learner-resource interactions from user-to-course and course-to-user perspectives, enhancing semantic alignment and improving matching accuracy between learner needs and instructional resources.

$$f_{\text{cross}} = \text{BiCross}(u, c) \tag{11}$$

where u denotes the student feature representation, c represents the educational resource or course feature representation, and fcross is the crossed feature representation generated by the Bidirectional Feature Cross (BiCross) module. The BiCross module captures higher-order interactions between student and educational resource fea-

tures through bidirectional feature crossing. Unlike simple feature concatenation, it models nonlinear dependencies across modalities, enabling richer semantic representations. This enhances feature expressiveness and improves cross-modal matching and personalized recommendation performance. Prior to output prediction, the interest representation h_t^* and crossed feature representation f_{cross} are transformed into a common latent feature space with consistent dimensionality. This alignment step ensures compatibility between sequential learner-interest representations and bidirectional feature interaction outputs, enabling effective feature fusion while preserving complementary semantic information for subsequent cross-modal matching and probability estimation.

Collaborative deep semantic learning framework

For unlabeled multimodal samples $X_u = \{x_1, x_2, \ldots, x_n\}$, embeddings v_i form $V_u \in R^{n\times k}$.

$$V_u = \left[v_1^\top; v_2^\top; \ldots; v_n^\top\right] \tag{12}$$

We normalize V_u row-wise into probability distributions over semantic categories:

$$\hat{S}_u(i,j) = \frac{\exp\left(v_{ij}\right)}{\sum_{j=1}^{k} \exp\left(v_{ij}\right)} \tag{13}$$

where v_{ij} denotes the semantic score of samples i for category j obtained from the semantic embedding matrix, $\hat{S}_u(i,j)$ represents the normalized probability that sample i belongs to semantic category j, and k denotes the total number of semantic categories. For each sample ii, we then assign its weak semantic label $\tilde{y}_i$ by selecting the category with the highest probability:

$$\tilde{y}_i = \operatorname{argmax} \tilde{S}_u(i,j) \tag{14}$$

where $\tilde{y}_i$ denotes the weak semantic label assigned to sample i, determined by selecting the semantic category with the highest probability in $\tilde{S}_u(i,j)$. Weak labels $\tilde{y}_i$ are iteratively fed back as pseudo-labeled data to align with the latent semantic graph improving cross-modal retrieval accuracy and robustness through global semantic relationships.

Semantic associations among educational concepts enhance contextual recommendation by capturing latent relationships beyond explicit feature similarity, improving relevance, coherence, and personalized learning support. WSSL-RSNN combines weak semantic labeling, embedding learning, temporal weighting, and recurrent refinement, iteratively updating multimodal representations through pseudo-labels to support adaptive learning and robust cross-modal retrieval beyond fixed-label approaches.

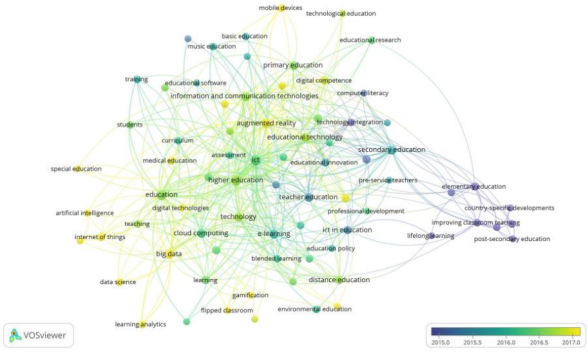

Fig. 6. Evolving Co-occurrence Network of Educational Technology and Moral Education

Fig. 6 illustrates temporal evolution of educational technology, innovation, and moral education concepts; node colors indicate publication year, enabling temporally weighted weak semantic label generation. $X_u = \{x_1, x_2, \ldots, x_n\}$, V_u is computed (Eq. 15), and $C \in R^{k\times k}$ is temporally weighted.

$$\hat{S}_u^*(i,j) = \frac{\tau_{j\cdot} \cdot \exp\left(v_{ij}\right)}{\sum_{j=1}^{k} \cdots \tau_{j\cdot} \cdot \exp\left(v_{ij}\right)} \tag{15}$$

where τ_j denotes the normalized temporal recency weight associated with semantic category j, v_{ij} represents the semantic score of sample i for category j, and $\hat{S}_u^*(i,j)$ denotes the temporally adjusted probability of assigning sample i to semantic category j. Within this formulation, the temporal weighting vector τ prioritizes semantically relevant categories according to their recency. Higher τ values increase category selection probability, improving weak label assignment and temporal sensitivity. We then assign the weak semantic label $\tilde{y}_i$ as:

$$\tilde{y}_i = \operatorname{argmax} \tilde{S}_j^*(i,j) \tag{16}$$

Pseudo-label assignment uses temporally adjusted category probabilities to suppress unreliable activations and prioritize consistently supported categories, improving the robustness, stability, and reliability of weak semantic label generation. For instance, when recent learner interactions increasingly emphasize entrepreneurial innovation compared with traditional management topics, the corresponding semantic category receives a higher temporal weight. Labels are adaptively assigned using semantic and temporal dynamics, improving recommendation refinement and crossmodal retrieval robustness in education.

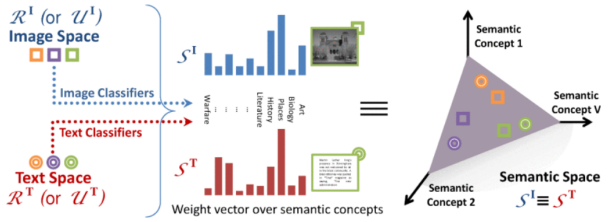

Fig. 7. The semantic matching mechanism

Semantic Matching Mechanism

Fig. 7 illustrates semantic matching, where image features R^I (U^I) and text features R^T (U^T) are projected into a shared V-dimensional space. Classifiers generate S^I and S^T, enabling effective cross-modal retrieval. These Vectors S^I and S^T lie in a shared V-dimensional semantic space, ensuring semantically similar images and texts map to nearby points for effective cross-modal retrieval. Initial embeddings from pre-trained features provide informative optimization starting points. Domain-specific adaptation with educational data improves convergence stability and captures modality-specific and cross-modal relationships more effectively. Formally, the projection functions f_I and f_T for images and texts are defined as:

$$S^I = f_I\left(x^I\right) \in R^V, \quad S^T = f_T\left(x^T\right) \in R^V \tag{17}$$

where x^I and x^T are the original image and text features, respectively. Image and text features are projected into a shared semantic embedding space of dimensionality V. Using L2-normalized embeddings and cosine similarity, modality-specific variations are reduced, preserving semantic relationships by clustering related image-text pairs while separating unrelated pairs. Semantic matching is then performed by computing the similarity between S^I and S^T in the semantic space. A common similarity metric such as cosine similarity is used:

$$\text{sim}\left(x^I, x^T\right) = \frac{\left\langle S^I, S^T \right\rangle}{\left\|S^I\right\|_2 \cdot \left\|S^T\right\|_2} \tag{18}$$

where $\langle \cdot, \cdot \rangle$ denotes the inner product. Cosine similarity is preferred over Euclidean distance because it captures angular relationships between semantic representations and is less sensitive to feature magnitude differences, making it suitable for multimodal retrieval with differently scaled textual and visual embeddings. By aligning the projections of both modalities into this shared semantic space, the model achieves effective and interpretable cross-modal retrieval. The shared semantic space projects textual and visual educational resources into a unified representation, reducing modality gaps and semantic mismatches, improving retrieval consistency, enabling reliable similarity evaluation, and supporting coherent multimodal content retrieval and recommendation.

To improve robustness, the model can also incorporate thresholding, where only candidates with similarity scores above a certain threshold θ are considered relevant:

$$\text{Relevant}\left(S^q, S^c\right) = \{1, \text{ if }, \text{sim}\left(S^q, S^c\right) \geq \theta 0, \text{ otherwise} \tag{19}$$

The cosine similarity threshold θ was optimized via validation to maximize mAP. Lower $\theta \backslash$ theta θ improves recall but may introduce irrelevant results, whereas higher $\theta \backslash$ theta θ enhances precision at the expense of recall, affecting ranking sensitivity. Threshold-based filtering retains candidates exceeding θ, improving recommendation precision and delivering more contextually relevant multimedia educational resources. This retrieval strategy effectively exploits the shared semantic space and ensures that the returned results align with the latent semantic concepts shared across modalities. The optimization strategy employs a composite objective comprising semantic matching and weak semantic supervision losses. L2 regularization and dropout are incorporated to mitigate overfitting, while normalized modality-specific gradients balance image and text contributions during backpropagation, ensuring stable multimodal representation learning and robust retrieval performance.

4. Result and discussion

Experiments and evaluation

This section evaluates the proposed platform and cross-modal retrieval model, validating semantic retrieval and pedagogical effectiveness.

Datasets

We evaluated benchmark datasets: Wikipedia (2,866 image-text pairs, 10 categories), NUS-WIDE (269,648 tagged images), and Wikipedia-CNN (CNN-based features), validating crossmodal semantic retrieval in controlled settings.

Innovation & entrepreneurship education dataset (i&e-edu):

A domain-specific dataset from a university course was used to assess real-world effectiveness, containing 5,200 multimedia materials, 1,200 multimodal submissions from 3 classes over 2 semesters, and 8 semantic categories. The training sample distribution exhibits moderate semantic-category imbalance, with some categories more frequent in

Table 1. Summary of Key Mathematical Notations

Symbol	Description
c_i^*	Optimal recommended course for student i
C	Set of available candidate courses
x_i	Feature vector of student i
P_i	Personalized learning path of student i
$U(c, x_i, P_i)$	Utility score function
h_i	Hidden state generated by the GRU layer
α_i	Attention weight associated with hidden state h_i
q	Learnable query vector used in the attention mechanism
v_{ij}	Semantic score of samples i for category j
$S^{\wedge}{}_u(i,j)$	Normalized probability that sample i belongs to semantic category j
$\tilde{y}_i$	Weak semantic label assigned to sample i
τ_j	Temporal recency weight associated with semantic category j
$S_u^{**}(i,j)$	Temporally adjusted probability of assigning sample i to semantic category j

Table 2. mAP (%) of MRCR-SNN and WSSL-RSNN on Cross-Modal Benchmarks

Dataset	Task	MRCR-SNN	WSSL-RSNN
Wikipedia	Image → Text	28.12	30.56
Wikipedia	Text → Image	20.19	21.97
Wikipedia-CNN	Average	30.96	33.14
NUS-WIDE	Average	29.25	30.72

innovation and entrepreneurship activities. WSSL-RSNN addresses this through weak semantic label generation and semantic-aware multimodal supervision, improving minority-category representation, reducing dominant-class bias, and enhancing generalization across diverse educational samples. All datasets were split into 70% training, 10% validation, and 20% testing.

Baselines

We compared WSSL-RSNN with CCA, PLS, BLM, MRCR-NN, and MRCR-SNN baselines, none of which incorporate weak semantic label generation or temporal weighting.

5. Results and analysis

Benchmark results

Table 2 shows WSSL-RSNN outperforming MRCR-SNN: Wikipedia 28.12→30.56, 20.19→21.97; Wikipedia-CNN 30.96→33.14; NUS-WIDE 29.25→30.72, validating weak semantic labels and temporal weighting.

Innovation & entrepreneurship application

As shown in Table 3, WSSL-RSNN surpassed MRCR-SNN with mAP 32.7% (+4.4%), P@10 49.4%(+6.3%), and 26.5% higher student engagement, improving semantic matching and ethical awareness.

Fig. 8 shows a radar chart (max 0.4) where CRM-SLIE outperforms NCF, Deep&Cross, GRU4Rec, STAMP, and LightSANs, especially on dimensions 2, 7, 9, and 10.

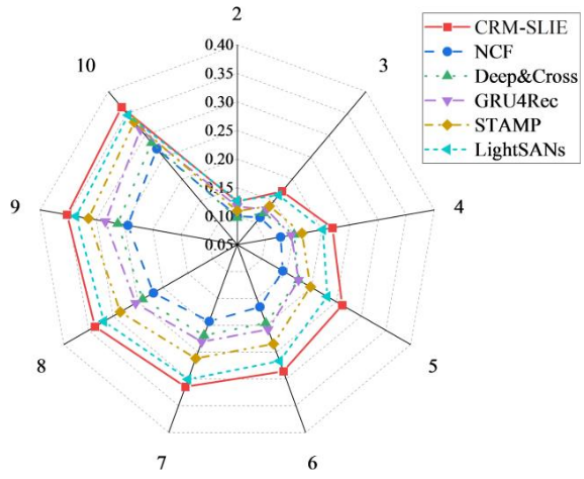

Fig. 8. Performance comparison of the proposed CRM-SLIE model and baseline methods

Fig. 9 shows six ablations; the full model reaches AUC ≈ 0.8725 and recall@ 10 ≈ 0.37, while No Feature Cross performs worst (0.8695, 0.28).

6. Conclusion

This paper presented an integrated teaching platform for higher education that combines innovation, entrepreneurship, and moral education within a multimedia-enabled, AI-driven framework. Addressing the limitations of ex-

Table 3. Performance of WSSL-RSNN vs. MRCR-SNN on I&E-EDU dataset

Metric	MRCR-SNN	WSSL-RSNN	Improvement
Mean Average Precision (mAP)	28.3%	32.7%	+4.4%
Precision@10 (P@10)	43.1%	49.4%	+6.3%
Student Engagement Increase	-	26.5%	N/A
Teacher Topic Alignment Score	Baseline	Higher	Qualitative
Critical Thinking Feedback	Baseline	Stronger	Qualitative

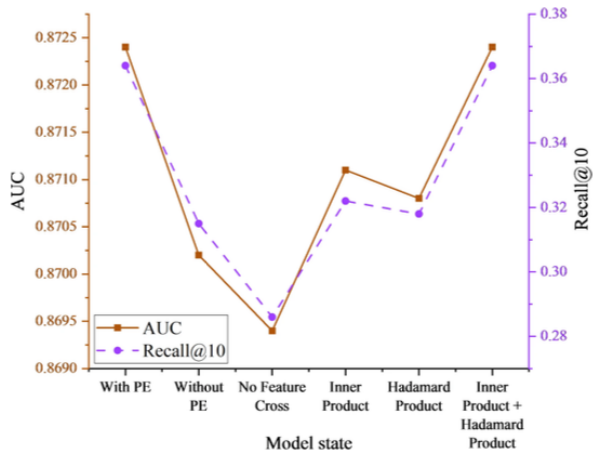

Fig. 9. AUC and Recall@10 under different configurations

isting IEE systems, which often lack personalization and neglect moral dimensions, the proposed platform leverages a hybrid neural network-based cross-modal retrieval model to align multimodal instructional resources and student learning behaviors in a shared semantic space. By introducing weak semantic label generation, bidirectional feature interaction, attention mechanisms, and position encoding, the model effectively captures the complex, dynamic relationships between diverse educational content and heterogeneous student profiles. Extensive quantitative and qualitative experiments demonstrate the platform's superior performance over baseline models in retrieval accuracy, recommendation quality, and student engagement. Future research will focus on extending the model to support additional modalities such as audio and video, incorporating explainable AI to enhance transparency and trust, and conducting large-scale field studies to assess long-term learning outcomes and behavioral impacts.

Acknowledgments

Not Applicable

Declarations

Acknowledgements: The authors would like to show sincere thanks to those techniques who have contributed to this research.

Consent for publication:

All authors reviewed the results, approved the final version of the manuscript and agreed to publish it.

Data availability:

The experimental data used to support the findings of this study are available from the corresponding author upon request.

Author contributions:

Zijin Li: Conceptualization, Methodology, Software, Data curation, Formal analysis, Visualization, Writing - original draft. Weijie Zhao: Supervision, Validation, Resources, Project administration, Writing - review & editing, All authors have read and approved the final manuscript.

Consent to participate:

Not applicable.
Consent to Publication: All authors have provided consent for publication of this manuscript.

Conflicts of interest:

The authors declared that they have no conflicts of interest regarding this work.

Funding statement:

There is no specific funding to support this research.

References

[1] S. Ruan and K. Lu, (2025) *"Adaptive deep reinforcement learning for personalized learning pathways: A multimodal data-driven approach with real-time feedback optimization"* **Computers and Education: Artificial Intelligence 9**: 100463. DOI: 10.1016/j.caeai.2025.100463.

[2] Y. Zhou and H. Zhou, (2022) *"Research on the quality evaluation of innovation and entrepreneurship education of college students based on extenics"* **Procedia Computer Science *199***: 605–612. DOI: 10.1016/j.procs.2022.01.074.

[3] F. Makhmudov, A. Kultimuratov, and Y. I. Cho, (2024) *"Enhancing multimodal emotion recognition through attention mechanisms in BERT and CNN architectures"* **Applied Sciences *14***(10): 4199. DOI: 10.3390/app14104199.

[4] X. Ji, L. Sun, and K. Huang, (2025) *"The construction and implementation direction of personalized learning model based on multimodal data fusion in the context of intelligent education"* **Cognitive Systems Research *92***: 101379. DOI: 10.1016/j.cogsys.2025.101379.

[5] T. Shengju, F. Li, Z. Wang, and X. Zhaoyuan, (2025) *"Cross-modal adaptive reconstruction of open education resources"* **Scientific Reports *15***(1): 30838. DOI: 10.1038/s41598-025-15200-8.

[6] P. Cantillon, W. De Grave, and T. Dornan, (2022) *"The social construction of teacher and learner identities in medicine and surgery"* **Medical Education *56***(6): 614–624. DOI: 10.1111/medu.14727.

[7] K. Liu, F. Xue, D. Guo, L. Wu, S. Li, and R. Hong, (2023) *"MEGCF: Multimodal entity graph collaborative filtering for personalized recommendation"* **ACM Transactions on Information Systems *41***(2): 1–27. DOI: 10.1145/3544106.

[8] I. Chetoui, E. El Bachari, and M. El Adnani, (2026) *"Multi-modal graph neural networks for cross-domain educational recommendation: integrating behavioral analytics and institutional context for personalized learning"* **Smart Learning Environments *13***(1): 26. DOI: 10.1186/s40561-026-00452-2.

[9] X. Xiao, (2025) *"MMAgentRec, a personalized multimodal recommendation agent with large language model"* **Scientific Reports *15***(1): 12062. DOI: 10.1038/s41598-025-96458-w.

[10] D. Maier, (2022) *"The use of wood waste from construction and demolition to produce sustainable bioenergy: a bibliometric review of the literature"* **International Journal of Energy Research *46***(9): 11640–11658. DOI: 10.1002/er.8021.

[11] V. Y. Dobrova, O. S. Popov, O. V. Shtrimaitis, O. O. Andreeva, and O. M. Proskurnia, (2022) *"Joint task force core competency framework adoption process at a national level: a survey of Ukrainian-based clinical research professionals"* **Therapeutic Innovation and Regulatory Science *56***(5): 814–821. DOI: 10.1007/s43441-022-00428-7.

[12] Y. Xu, W. Li, J. Tai, and C. Zhang, (2022) *"A bibliometric-based analytical framework for the study of smart city lifeforms in China"* **International Journal of Environmental Research and Public Health *19***(22): 14762. DOI: 10.3390/ijerph192214762.

[13] R. Mavilia and R. Pisani, (2022) *"Blockchain for agricultural sector: the case of South Africa"* **African Journal of Science, Technology, Innovation and Development *14***(3): 845–851. DOI: 10.1080/20421338.2021.1908660.

[14] B. C. Lines, R. Kakarapalli, and P. H. D. Nguyen, (2022) *"Does best value procurement cost more than low-bid? A total project cost perspective"* **International Journal of Construction Education and Research *18***(1): 85–100. DOI: 10.1080/15578771.2020.1777489.

[15] J. Liu, E. Gong, and X. Wang, (2022) *"Economic benefits of construction waste recycling enterprises under tax incentive policies"* **Environmental Science and Pollution Research *29***(9): 12574–12588. DOI: 10.1007/s11356-021-13831-8.