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

Traitement du Signal
Vol. 43, No. 2, April, 2026, pp. 821-837
Journal homepage: http://iieta.org/journals/ts

Optimizing Multimodal Image Representation Learning with Knowledge Graph Semantic Constraints for Personalized Education

Yalin Li[1*], Wei Zhang[2], Liangbo Zhang[3]

[1] School of Environmental and Biological Engineering, Henan University of Engineering, Zhengzhou 451191, China
[2] School of Ecology and Environment, Zhengzhou University, Zhengzhou 450001, China
[3] School of Environmental Engineering, Henan University of Technology, Zhengzhou 450001, China

Corresponding Author Email: li_ya_lin@outlook.com

https://doi.org/10.18280/ts.430222

Received: 11 December 2025
Revised: 12 February 2026
Accepted: 9 March 2026
Available online: 30 April 2026

Keywords:
multimodal image representation, knowledge graph, sparse graph attention, contrastive learning, cognitive state, hyperbolic geometry, personalized education

ABSTRACT

In personalized education scenarios, visual representations of instructional images often suffer from a lack of high-level semantics, the absence of knowledge-driven logical constraints in Transformer self-attention mechanisms, and an inability to adaptively modulate features according to students' cognitive states. These limitations significantly hinder the application of multimodal image understanding in personalized learning services. To address these challenges, this paper proposes an optimized multimodal image representation learning framework guided by knowledge graph semantic constraints. The proposed method begins with a multimodal feature extraction and semantic prototype initialization module, which maps entities and relations from educational knowledge graphs into the visual feature space. A semantic-guided sparse graph attention mechanism is then introduced to select semantically relevant image regions based on knowledge topology, thereby reducing computational complexity while explicitly injecting structured semantic constraints. A Relation-Path-Constrained (RPC) - Information Noise-Contrastive Estimation (InfoNCE) contrastive loss is further designed to encode causal relationships and partial-order evolutionary patterns among educational concepts. In addition, student cognitive state vectors are incorporated to enable dynamic semantic masking and personalized weighting of image representations at the attention level. To further refine the representation distribution, the framework integrates graph Laplacian semantic consistency regularization and hyperbolic geometric hierarchical constraints, optimizing the learned embeddings from both topological alignment and hierarchical modeling perspectives. The entire model is trained in an end-to-end manner. Extensive experiments on a self-constructed educational multimodal image dataset and public benchmarks demonstrate that the proposed method consistently outperforms strong baselines across multiple vision-centric tasks, including cross-modal image retrieval, knowledge point clustering, and personalized exercise recommendation. Visualization analyses—via attention heatmaps and hyperbolic space embeddings—further confirm the model's precise semantic focus, effective hierarchical modeling capability, and adaptive personalization characteristics. This study offers a novel technical paradigm for image processing and multimodal understanding in personalized education, with significant academic value and practical potential.

1. INTRODUCTION

The in-depth development of digital education has propelled personalized intelligent education into a refined stage [1-3]. Instructional images, such as textbook diagrams, physics and chemistry experiment illustrations, and geometric structure charts, have become core visual carriers for transmitting knowledge and assisting cognition. As key technologies for parsing the semantic content of instructional images, mining the intrinsic associations between knowledge points, and enabling the precise delivery of personalized learning content [4, 5], the performance of image processing and multimodal representation learning directly determines the quality and efficiency of personalized education services. Traditional image processing methods [6] are limited to extracting low-level pixel, texture, and contour features, making it difficult to capture the educational knowledge points, conceptual hierarchies, and causal logical relationships embedded within instructional images. Although general-purpose visual multimodal models [7] achieve cross-modal alignment between images and text, their design lacks domain-specific prior knowledge from education. They suffer from inherent defects such as global redundancy in attention, ambiguous semantic associations, and susceptibility to background noise, failing to meet the core requirements of personalized education for knowledge-structured, cognitively adaptive, and semantically interpretable image representations. This has become a critical bottleneck restricting the application of multimodal technologies in personalized education scenarios [8]. From an academic research perspective, this paper establishes a deep integration paradigm between knowledge graph topological constraints and visual representation learning [9, 10], breaking through the limitation of general multimodal models lacking domain knowledge guidance, and providing new ideas for domain-specific instructional image multimodal representation modeling. From an engineering application perspective, the proposed method can directly support downstream image processing tasks such as knowledge point localization in teaching images, cross-modal retrieval, visual diagnosis of learning weaknesses, and personalized exercise recommendation [11, 12], demonstrating significant theoretical research value and engineering implementation potential.

Existing related research still suffers from many shortcomings that need to be addressed urgently, making it difficult to adapt to the special needs of personalized education scenarios. In the field of general multimodal visual representation, Vision Transformer (ViT), Contrastive Language Image Pretraining (CLIP), and various derivative cross-modal models [13, 14] primarily rely on global self-attention mechanisms for feature interaction. They not only suffer from high computational complexity of $O(N^2)$, but also lack effective semantic prior guidance. They can only achieve shallow image-text matching and fail to accurately model the hierarchical relationships and causal evolution paths of knowledge points within instructional images. Regarding knowledge graph-enabled visual modeling [15, 16], existing studies mostly stop at the simple concatenation or shallow mapping of knowledge entity embeddings and visual features. They fail to construct knowledge topology-driven sparse visual attention mechanisms and lack relation path-level contrastive constraints, making it difficult to deeply integrate the structured semantics of knowledge graphs into the entire process of image representation learning, resulting in insufficient semantic logic in the representations [17, 18]. In the field of personalized education modeling, existing research mostly focuses on knowledge tracing modeling based on text and behavioral data [19, 20], rarely introducing student cognitive states from the perspective of image processing, and thus unable to achieve dynamic adaptive modulation of instructional image region representations according to students' knowledge mastery levels [21, 22]. Meanwhile, existing models lack effective representation optimization means such as graph topological regularization and hyperbolic hierarchical geometric constraints, leading to significant limitations in the generalization and semantic interpretability of the models.

Aiming at the aforementioned research gaps, this paper conducts research on optimizing multimodal image representation learning for personalized education teaching image processing scenarios. The main contributions are as follows: (1) A complete framework for multimodal image representation learning integrating knowledge graph semantic constraints is proposed. A semantically guided sparse graph attention mechanism is innovatively designed to filter interaction relationships between image regions based on the topological structure of knowledge concepts. This significantly reduces the computational overhead of traditional self-attention while achieving structured constraints of high-level knowledge semantics on visual attention, thereby enhancing the semantic focusing capability of the representation. (2) A Relation-Path-Constrained (RPC) - Information Noise-Contrastive Estimation (InfoNCE) contrastive learning loss is constructed. It breaks through the limitation of traditional InfoNCE loss, which can only achieve sample-level image-text alignment, by explicitly modeling the causal partial order and evolutionary relationships of educational knowledge points, thereby strengthening the logical semantic encoding capability of multimodal image representations. (3) A cognitive state-driven dynamic semantic mask optimization strategy is proposed. By taking students' knowledge mastery as prior input, the weights of image regions are adaptively modulated at the visual attention aggregation layer. For the first time, personalized customization of instructional image representations is achieved from the perspective of image processing attention mechanisms, adapting to the cognitive differences of different students. (4) Graph Laplacian semantic consistency regularization and hyperbolic geometric hierarchical constraints are introduced to optimize the distribution of image representations from the dimensions of knowledge topology alignment and educational concept hierarchy modeling, respectively. This effectively improves the semantic consistency, hierarchical modeling capability, and model interpretability of the representations, realizing end-to-end joint optimization training of the entire framework.

The organization of the subsequent chapters of this paper is as follows: Chapter 2 systematically reviews three types of related research: multimodal visual representation, knowledge graph and visual semantic fusion, and cognition-driven personalized education modeling, and deeply analyzes the inherent defects and shortcomings of existing research. Chapter 3 elaborates on the overall architecture of the optimization method proposed in this paper, the technical details of each core module, and the derivation of relevant mathematical formulas. Chapter 4 designs multiple groups of comparative experiments, ablation experiments, parameter sensitivity experiments, visualization experiments, and generalization experiments, verifying the effectiveness and superiority of the proposed method through quantitative results and qualitative analysis. Chapter 5 deeply analyzes the internal working mechanism of the proposed method, its core differences from existing methods, the limitations of the model, and directions for future research expansion. Chapter 6 summarizes the full text's research work and core conclusions, condensing the academic value and application prospects of the research findings.

2. OPTIMIZED MULTIMODAL IMAGE REPRESENTATION LEARNING METHOD INTEGRATING KNOWLEDGE GRAPH SEMANTIC CONSTRAINTS

2.1 Overview of overall architecture

To meet the core requirements of personalized education scenarios for semantic enhancement, cognitive adaptation, and interpretability of instructional image representations, this paper proposes an optimized multimodal image representation learning method integrating knowledge graph semantic constraints. Its core objective is to construct image representations capable of accurately capturing the semantics of educational knowledge points and adapting to the cognitive differences of students. The overall framework is illustrated in Figure 1. The framework consists of three collaboratively linked core modules, forming an end-to-end trainable multimodal representation learning system: (1) The Multimodal Feature Extraction and Initial Alignment Module extracts initial features of instructional image patches and relevant textbook text via a ViT and a lightweight text encoder, respectively, achieving preliminary spatial alignment of features from the two modalities. (2) The Knowledge Graph Semantic Constraint Module encodes entities and relations from the educational domain knowledge graph into a differentiable semantic topology. Relying on a sparse graph attention mechanism and relational path contrastive learning, it imposes structured semantic constraints on visual features, strengthening the semantic logic and accuracy of the representations. (3) The Personalized Representation Optimization Module dynamically generates semantic masks based on student cognitive state vectors, adaptively modulating the representation weights of image regions to realize the personalized adaptation of image representations. The following sections will elaborate on the core technical details and mathematical formalizations of each module in turn, clarifying the collaborative working mechanisms and operational principles of each module. The overall architecture is shown in Figure 1.

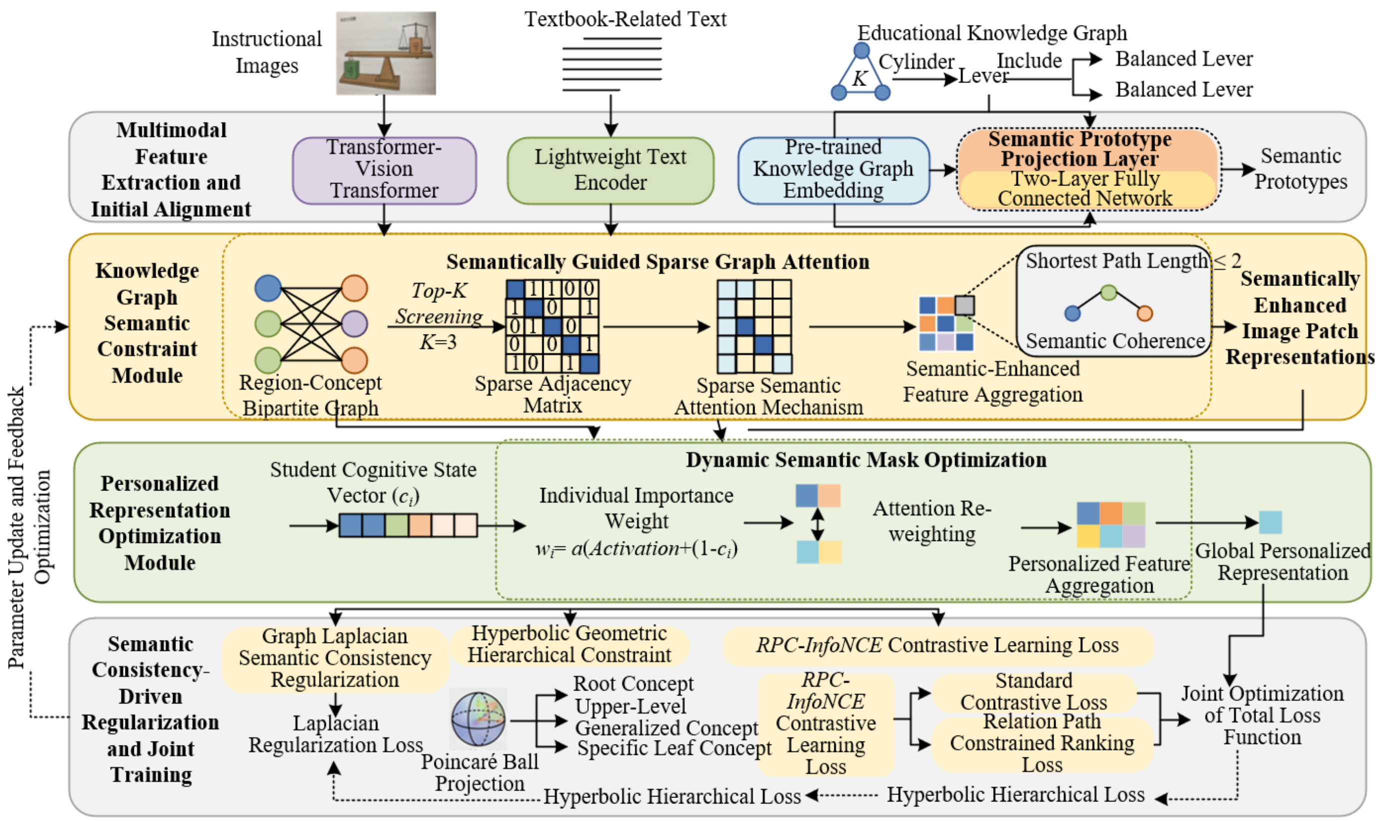


Figure 1. Overall architecture diagram of multimodal image representation learning integrating knowledge graph semantic constraints

2.2 Multimodal feature extraction and semantic prototype initialization

The core of multimodal feature extraction is to obtain discriminative initial features of images and text and achieve preliminary spatial alignment between the two modalities, laying the foundation for subsequent semantic constraints and personalized optimization. Given an instructional image $I \in R^{H\times W\times 3}$, where H and W are the height and width of the image respectively, and 3 represents the RGB channels, it is divided into N non-overlapping 16 × 16 image patches. The selection of this patch size balances the local details and global correlations of knowledge points in instructional images, effectively preserving key visual information in scenarios such as textbook diagrams and experiment illustrations. Each image patch undergoes linear projection processing to convert features from the pixel space into high-dimensional feature embeddings. The linear projection process can be expressed as:

$$v_j = W_{proj} \cdot x_j + b_{proj} \tag{1}$$

where, $x_j \in R^{16\times16\times3}$ is the pixel matrix of the j-th image patch, $W_{proj} \in R^{D\times768}$ and $b_{proj} \in R^D$ are the learnable weight and bias of the linear projection layer respectively, and D is the feature embedding dimension. Finally, the image patch embedding sequence $V=[v_1,v_2,\cdots,v_N]\in R^{N\times D}$ is obtained. To improve feature stability, layer normalization is introduced after projection to standardize each image patch embedding, reducing interference caused by feature distribution differences.

The structured semantics contained in the knowledge graph are key to improving the semantic logic of image representations. However, the modality gap between entity embeddings and visual features prevents the effective fusion of semantic information. Therefore, it is necessary to construct a semantic prototype projection mechanism to achieve spatial alignment between the two. The entity set $E = \{e_1,\ldots,e_M\}$ and relation triples (e_h,r,e_t) related to the current learning content are extracted from the educational knowledge graph, where M is the number of entities, and e_h、e_t are the head and tail entities of relation r, respectively. Each entity e_i obtains an initial vector $e_i \in R^D$ through a pre-trained knowledge graph embedding method. On this basis, a learnable semantic prototype projection layer is introduced to map the entity embeddings to the visual feature space, generating semantic prototypes with the same dimension as the image features. This projection layer adopts a two-layer fully connected network structure, which can fully fit the complex mapping relationship between entity semantics and visual features. Its mathematical expression is:

$$p_i = W_2 \cdot ReLU(W_1 \cdot e_i + b_1) + b_2 \tag{2}$$

where, $W_1 \in R^{D\times D}$ are $W_2 \in \mathrm{R}^{D\times D}$ the weights of the fully connected layers, $b_1 \in R^D$ and $b_2 \in R^D$ are the bias terms, and *ReLU* is the activation function used to enhance the nonlinear fitting capability of the model. p_i is the semantic prototype of the i-th concept. The semantic prototype projection layer is jointly optimized with the overall network. During the training process, its parameters are updated synchronously with those of subsequent modules such as the attention mechanism and contrastive learning, ensuring that the semantic prototypes can accurately adapt to the visual feature space. This achieves deep alignment between knowledge graph entity semantics and image visual features, providing a reliable semantic benchmark for the subsequent semantic constraint module.

Semantic prototypes are not only the carrier of knowledge graph semantics mapped to the visual space but will also serve as the core reference for the subsequent semantically guided sparse graph attention mechanism, used to filter semantically related image regions. Through the above process of multimodal feature extraction and semantic prototype initialization, the discriminability of the initial image and text features is guaranteed, and a bridge between knowledge semantics and visual features is built through semantic prototypes. This effectively eliminates fusion barriers caused by modality differences, providing a solid foundation for semantic enhancement and personalized optimization of the entire multimodal image representation learning framework.

2.3 Semantically guided sparse graph attention mechanism

The visual attention mechanism is the core of multimodal image representation learning. Traditional self-attention mechanisms achieve feature interaction by calculating the affinity between all pairs of image patches, which not only suffers from high computational complexity of $O(N^2)$, but also lacks effective semantic prior guidance. This easily leads to attention being focused on irrelevant background regions, making it difficult to adapt to the characteristics of instructional images that are knowledge-intensive and strongly semantically structured. To address this issue, a semantically guided sparse graph attention mechanism is constructed. Relying on the structured semantics of the educational knowledge graph, it filters semantically related image regions for attention interaction. While reducing computational complexity, it achieves precise constraints of high-level knowledge semantics on visual attention, injecting structured semantic information into the image representations. This mechanism takes the semantic prototypes generated in Section 2.2 as the core reference, and establishes the association between image patches and knowledge concepts by constructing a region-concept bipartite graph, thereby realizing the sparsification and semanticization of attention interaction. Figure 2 illustrates the schematic diagram of the semantically guided sparse graph attention interaction and feature aggregation mechanism.

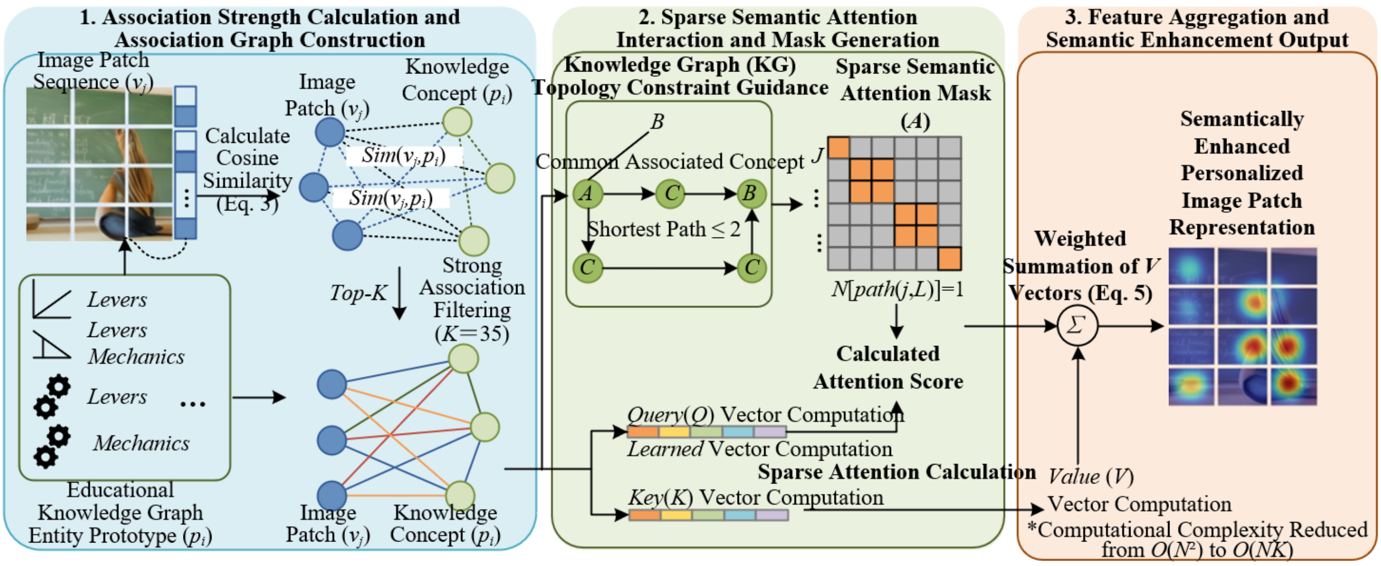


Figure 2. Schematic diagram of semantically guided sparse graph attention interaction and feature aggregation mechanism

The association strength between image patches and knowledge concepts is the core basis for sparse attention filtering, and it is necessary to quantify the semantic matching degree between the two through similarity calculation. For each image patch feature v_j and each concept semantic prototype p_i, cosine similarity is adopted to measure their semantic association degree. This metric can effectively avoid interference caused by feature scale differences and accurately capture the semantic consistency between the two. The calculation method is:

$$s_{j,i} = \frac{v_j \cdot p_i}{\|v_j\| \|p_i\|} \tag{3}$$

where, $s_{j,i} \in [0,1]$, and a larger value indicates a closer semantic association between the j-th image patch and the i-th knowledge concept. To ensure the sparsity and semantic specificity of attention interaction, for each image patch j, the top K knowledge concepts with the highest similarity are selected as its strongly associated concepts. The range of K is set to 3 to 5, which can not only guarantee the semantic integrity of the image patch but also avoid semantic redundancy caused by introducing too many concepts. Based on this, a sparse adjacency matrix $A \in \mathbb{R}^{(N \times M)}$ is constructed, where $A_{j,i} = 1$ indicates that the i-th concept is a *Top-K* strongly associated concept of the j-th image patch, otherwise $A_{j,i} = 0$. The sparsity of the adjacency matrix is determined by the value of K. Through this matrix, the knowledge concepts corresponding to each image patch can be clearly marked, providing a clear basis for the screening of subsequent attention interactions.

Based on the image patch-concept associations marked by the adjacency matrix, a sparse semantic attention mask is further constructed to achieve precise screening of semantically related image regions and attention weight allocation. The calculation of attention scores needs to take into account both the interactive affinity of image patch features and the relevance of knowledge semantics. Its mathematical expression is:

$$\alpha_{j,l} = \frac{exp\left(\left(W_Q v_j\right)^{\top}(W_K v_l)/\sqrt{d}\right) \cdot I[path(j,l)]}{\sum_{m=1}^{N} exp\left(\left(W_Q v_j\right)^{\top}(W_K v_m)/\sqrt{d}\right) \cdot I[path(j,m)]} \tag{4}$$

where, $W_Q \in \mathbb{R}^{(D \times d)}$ and $W_K \in \mathbb{R}^{(D \times d)}$ are the learnable weight matrices for the query and key vectors, respectively; d is the feature dimension of the attention head, usually satisfying $D = head \times d$ (where $head$ is the number of attention heads); $\sqrt{d}$ is used to alleviate the gradient vanishing problem during the attention score calculation process; $\mathbb{N}[path(j,l)]$ is an indicator function whose value is determined by the semantic association relationship between image patche j and l. When the two share at least one common strongly associated concept, or the shortest path length between their strongly associated concepts in the knowledge graph is ≤2, $\mathbb{N}[path(j,l)] = 1$, otherwise it is 0. This design fully integrates the topological structure of the educational knowledge graph, ensuring that only image patches that are semantically directly or indirectly associated can perform attention interaction. For example, image regions belonging to the same "Mechanics" concept group, or regions indirectly associated through concepts such as "Fulcrum" and "Lever," effectively excluding interference from background noise and irrelevant regions. Meanwhile, the computational complexity of attention interaction is reduced from $O(N^2)$ to $O(NK)$, significantly improving the model training and inference efficiency.

After the attention weight calculation is completed, the semantic-enhanced aggregation of image patch features is realized through weighted summation, obtaining the semantically constrained image patch representation v_j'. Its calculation method is:

$$v_j' = \sum_{l=1}^{N} \alpha_{j,l} W_V v_l \tag{5}$$

where, $W_V \in \mathbb{R}^{(D \times d)}$ is the learnable weight matrix for the value vector, and $\alpha_{j,l}$ is the attention weight of the j-th image patch on the l-th image patch. A higher weight indicates a greater semantic contribution of that image patch to the current image patch. Through the above process, the semantically guided sparse graph attention mechanism not only realizes the sparsification and semanticization of attention interaction but also deeply integrates the structured semantics of the knowledge graph into the visual feature aggregation process, enabling image patch representations to accurately capture the semantic associations and logical relationships of knowledge points. The enhanced image patch representations generated by this mechanism will serve as the basis for subsequent contrastive learning and personalized optimization, providing strong support for improving the semantic logic and interpretability of multimodal image representations.

2.4 Contrastive learning under relational path constraints

Contrastive learning is a key means to achieve multimodal feature alignment and semantic enhancement. Its core lies in guiding the model to learn discriminative representations by constructing positive and negative sample pairs. Instructional images in educational scenarios often contain causal associations and partial-order evolutionary relationships among multiple knowledge points. Traditional contrastive learning methods can only achieve sample-level alignment between images and corresponding texts, failing to capture such structured semantic relationships. This leads to a lack of knowledge logic association in image representations, making it difficult to meet the requirements of personalized education for knowledge point semantic modeling. To this end, a contrastive learning mechanism under relational path constraints is constructed, integrating the relational path information from the knowledge graph into the design of the contrastive learning loss. This enables image representations to accurately encode the causal and partial-order relationships of educational knowledge points, further enhancing the semantic logic and discriminative ability of the representations. Figure 3 illustrates the framework diagram of the RPC-InfoNCE contrastive learning model under relational path constraints.

The accurate acquisition of image global representations and text representations is the foundation of contrastive learning. It is necessary to construct a unified dimensional multimodal representation space based on the semantically enhanced image patch representations and knowledge graph entity texts mentioned above. For the image patch representation processed by the semantically guided sparse graph attention mechanism v_j', the global average pooling operation is adopted to extract the image global representation z_I. This operation can effectively aggregate the semantic information of all knowledge point regions in the image, balancing local details and global correlations. Its mathematical expression is:

$$z_I = \frac{1}{N} \sum_{j=1}^{N} v_j' \tag{6}$$

where, N is the number of image patches, and $z_I \in R^D$ is the image global representation. For each entity in the knowledge graph related to the image e, text fragments such as concept definitions and knowledge point descriptions are extracted from the textbook, and feature extraction is performed through a lightweight text encoder to obtain the entity text representation $t_e \in \mathrm{R}^{\mathrm{D}}$. This ensures that the text representation and the image representation are in the same feature space, providing a basis for subsequent contrastive learning. Meanwhile, the relational path set P related to the current image is extracted from the knowledge graph. Each path corresponds to the causal evolution or partial-order relationship of knowledge points. The path length L is set to 3 to 5 according to the logical complexity of educational knowledge to ensure that the path can completely reflect the association logic between knowledge points.

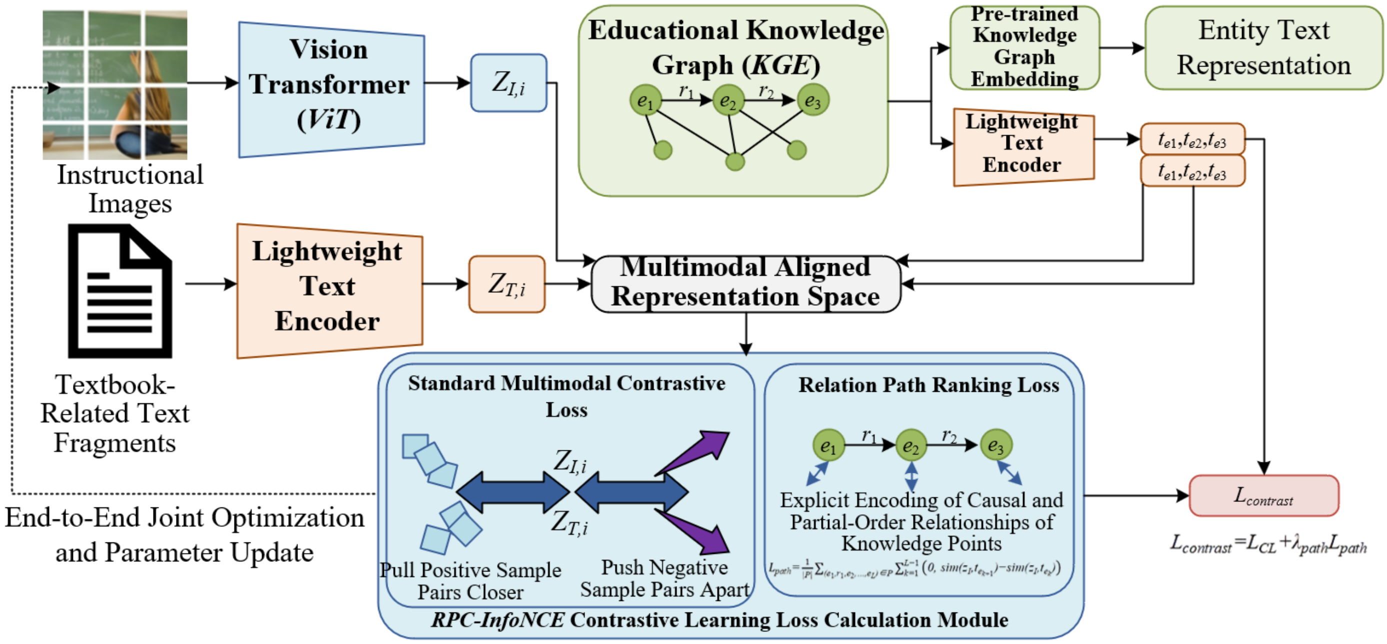


Figure 3. Framework diagram of the Relation-Path-Constrained (RPC)- Information Noise-Contrastive Estimation (InfoNCE) contrastive learning model under relational path constraints

To enable image representations to encode the relational path information in the knowledge graph, a path ranking loss is designed. By constraining the similarity relationship between the image global representation and the entity text representations along the path, explicit encoding of the causal and partial-order relationships of knowledge points is achieved. The core of the path ranking loss is to force the image representation to maintain a higher similarity with the text representations of subsequent entities along the path direction, thereby restoring the evolution order of knowledge points in the representation space. Its mathematical expression is:

$$L_{path}=\frac{1}{|P|}\sum_{(e_1,r_1,e_2,\dots,e_L)\in P}\sum_{k=1}^{L-1}\left(0,\ sim(z_I,t_{e_{k+1}})\text{-}sim(z_I,t_{e_k})\right) \quad (7)$$

where, $|P|$ is the number of valid relational paths in the training batch; γ is the margin parameter, ranging from 0.1 to 0.3, used to control the similarity difference between adjacent entity text representations and the image representation on the path, avoiding gradient vanishing or training instability; *sim* uses cosine similarity to quantify the semantic matching degree between the image global representation and the entity text representation. By introducing indicative constraints, this loss function produces a positive loss value when the similarity between the image representation and the subsequent entity text on the path is smaller than that with the preceding entity text. This forces the model to adjust parameters and optimize the image representation so that it can follow the causal evolution logic of knowledge points.

To ensure the basic alignment capability of multimodal features, the standard multimodal contrastive loss is retained, working synergistically with the path ranking loss to construct the complete total contrastive learning loss. The standard multimodal contrastive loss maximizes the similarity of positive sample pairs and minimizes the similarity of negative sample pairs, achieving sample-level alignment between images and corresponding texts. Its mathematical expression is:

$$L_{CL}=-\frac{1}{B}\sum_{i=1}^{B} log\frac{exp(sim(z_{I_i},t_{T_i})/\tau)}{\sum_{j=1}^{B} exp\,(sim(z_{I_i},t_{T_j})/\tau)} \quad (8)$$

where, B is the training batch size, z_{I_i} is the global representation of the i-th image, t_{T_i} is the representation of its corresponding text, and τ is the temperature parameter, ranging from 0.05 to 0.1, used to adjust the concentration of the similarity distribution and enhance the discriminative effect of contrastive learning. The total contrastive learning loss fuses the path ranking loss and the standard contrastive loss through the weight parameter λ_{path}, achieving the dual goals of basic image-text alignment and knowledge path constraints. Its expression is:

$$L_{contrast}=L_{CL}+\lambda_{path}L_{path} \quad (9)$$

where, λ_{path} is the weight coefficient of the path ranking loss, ranging from 0.5 to 1.0, used to balance the contributions of the two losses. This ensures that the model can not only achieve accurate multimodal alignment but also effectively encode the relational path information of knowledge points. This contrastive learning mechanism deeply integrates the structured semantics of the knowledge graph into the representation learning process, enabling image representations to possess good discriminability while accurately capturing the causal and partial-order relationships of educational knowledge points, providing a semantic foundation for subsequent personalized representation optimization.

2.5 Cognitive state-driven dynamic mask optimization

The core requirement of personalized education is to achieve precise adaptation between teaching content and students' cognitive levels. Reflected in image representation learning, this requires that the same instructional image generates differentiated representations for students with different levels of knowledge mastery. It should strengthen the image regions related to students' knowledge weaknesses while suppressing the regions corresponding to knowledge points already mastered. To achieve this goal, student cognitive state vectors are introduced as prior information into the image representation optimization process. A dynamic semantic mask mechanism is constructed to realize personalized modulation at the attention level, enabling image representations to adaptively match students' cognitive differences. The cognitive state vector can be output in real-time by a knowledge tracing model; this paper takes it as model input to ensure the timeliness and accuracy of cognitive information. Figure 4 presents the computational flow diagram of cognitive state-driven dynamic semantic masking and personalized feature generation.

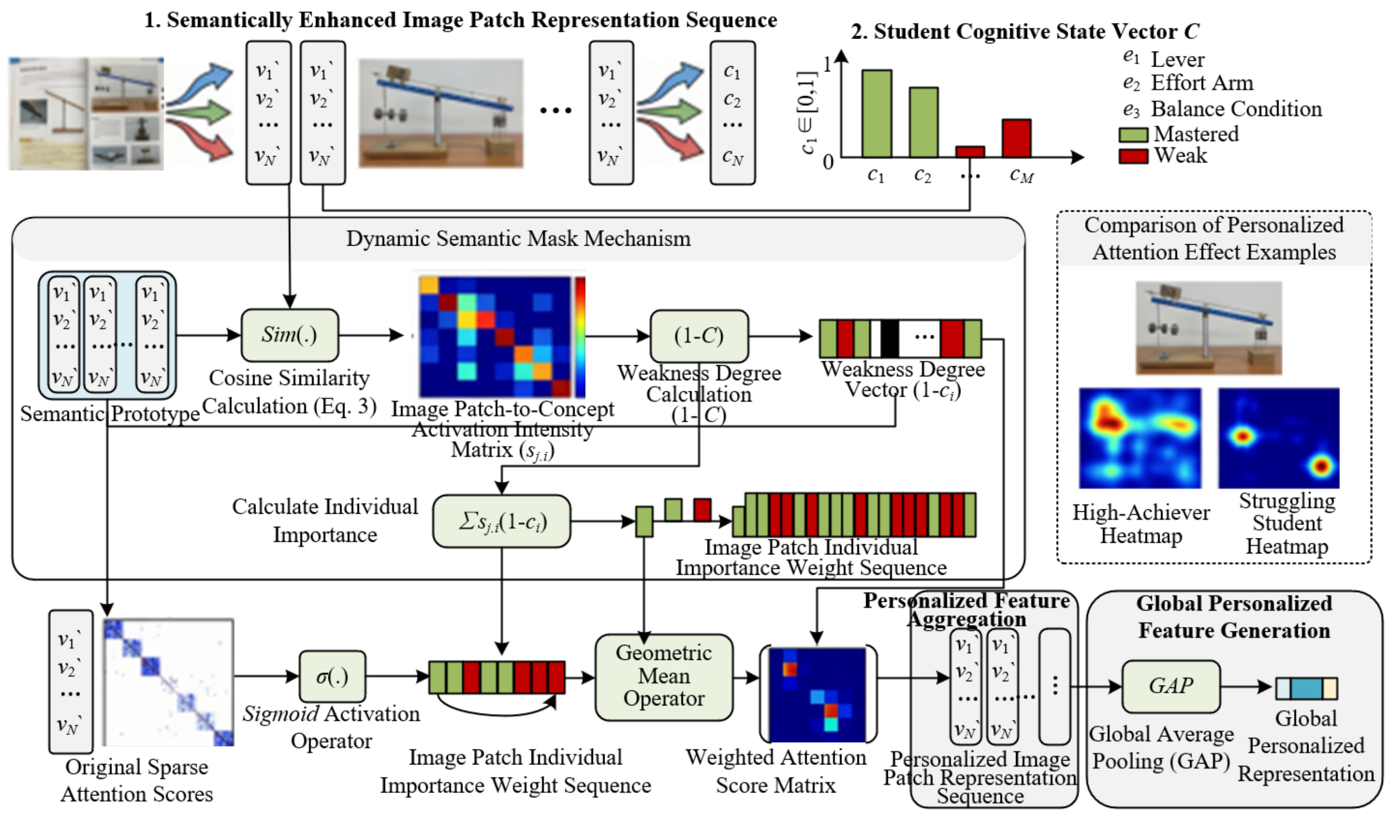


Figure 4. Computational flow diagram of cognitive state-driven dynamic semantic masking and personalized feature generation

The student cognitive state vector is used to quantify the student's mastery degree of each concept in the knowledge graph. It is defined as $c\in[0,1]^M$, where M is the total number of concepts in the knowledge graph, and c_i represents the probability that the student has mastered the i-th concept e_i. A value closer to 1 indicates a higher mastery level of the concept, while a value closer to 0 indicates that the concept is a knowledge weakness for the student. The generation of the dynamic semantic mask is based on the activation intensity between image patches and concepts, where the activation intensity $s_{j,i}$ of image patch j on concept i follows the cosine similarity calculated in Section 2.3, which can accurately reflect the semantic association degree between the image patch and the corresponding concept. Based on this, the individual importance weight of the image patch is defined to measure the learning value of the image patch for the current student. Its mathematical expression is:

$$w_j=\sigma\left(\sum_{i=1}^{M} s_{j,i}\cdot(1\text{-}c_i)\right) \tag{10}$$

where, $\sigma(\cdot)$ is the sigmoid activation function, specifically $\sigma(x)$=1/1+e^{-x}, used to normalize the weight value to the [0,1] interval to achieve a soft mask effect. From a physical perspective, if an image patch strongly activates a knowledge weakness concept (i.e., $s_{j,i}$ is large and c_i is small), then takes a larger value (1-c_i), causing w_j to approach 1, and the semantic information of that image patch will be preserved or even enhanced. If the image patch activates concepts that the student has already mastered, then w_j approaches 0, and the semantic information of that image patch will be softly suppressed. This design not only achieves personalized modulation but also avoids training instability and feature distortion caused by hard masks through the smoothing property of the sigmoid function.

To further improve training stability and avoid feature distortion caused by directly multiplying individual importance weights onto image patch features, the weights are integrated into the re-weighting process of attention scores, realizing the deep fusion of dynamic semantic masks and the attention mechanism. The re-weighted attention scores retain both the semantic-guided sparse characteristics from Section

2.3 and the personalized constraints of the student cognitive state. Its mathematical expression is:

$$\tilde{\alpha}_{j,l}=\frac{\alpha_{j,l}\cdot\sqrt{w_j w_l}}{\sum_{m=1}^{N}\alpha_{j,m}\cdot\sqrt{w_j w_m}} \tag{11}$$

where, $\alpha_{j,l}$ is the semantically guided sparse attention score obtained in Section 2.3, and $\sqrt{w_j w_m}$ is the geometric mean of the individual weights of image patches j and l, used to balance the contribution of the two image patches in attention interaction. This design allows the model to automatically reduce the attention weight from regions already mastered by the student during feature aggregation, while increasing the weight proportion of regions representing knowledge weaknesses, ensuring that the image representation focuses on the areas with the most learning value for the current student. The introduction of the geometric mean can effectively avoid extreme impacts caused by excessively high or low single image patch weights, ensuring the stability and rationality of the attention aggregation process.

Based on the re-weighted attention scores, feature aggregation is performed to obtain the personalized image patch representation. Its calculation method is:

$$\tilde{v}_j=\sum_{l=1}^{N}\tilde{\alpha}_{j,l}\,W_V v_l \tag{12}$$

where, W_V is the learnable weight matrix for the value vector, consistent with Section 2.3, to ensure consistency in the feature space. To obtain the global personalized image representation adapted to the student's cognitive state, global average pooling is performed on all personalized image patch representations $\tilde{v}_j$ to obtain the global personalized representation $\tilde{z}_I$. Its expression is:

$$\tilde{z}_I=\frac{1}{N}\sum_{j=1}^{N}\tilde{v}_j \tag{13}$$

This global representation fully integrates the structured semantics of the knowledge graph and the cognitive differences of students, accurately matching the knowledge mastery levels of different students, providing strong support for downstream tasks such as personalized exercise recommendation and learning weakness diagnosis. The entire dynamic mask optimization process is trained collaboratively with other modules of the model, ensuring the precision and effectiveness of personalized modulation, and truly realizing the personalized adaptation of instructional image representations.

2.6 Semantic consistency-driven regularization terms

To effectively prevent the model from overfitting to pseudo-semantic correlations and enhance the semantic consistency and interpretability of image representations, two targeted regularization terms are introduced relying on the topological structure of the knowledge graph. These terms optimize the representation distribution from the dimensions of knowledge topology alignment and concept hierarchy modeling, respectively. They ensure that the image representations learned by the model accurately match the structured characteristics and hierarchical logic of educational knowledge, collaborating with the semantic constraint and personalized optimization modules mentioned above to achieve end-to-end joint optimization.

The core of Graph Laplacian Semantic Consistency Regularization is to drive the similarity relationship between image patch representations to be consistent with the semantic distance of corresponding concepts in the knowledge graph, thereby strengthening the semantic rationality of the representations. First, the activated concept set corresponding to all image patches C_{active} is extracted, and an undirected graph is constructed with image patches as nodes. The edge weights in this graph are determined by the semantic distance between the concepts activated by the corresponding image patches in the knowledge graph. The edge weight calculation adopts an exponential decay function, which can transform the semantic distance between concepts into discriminative weight values. Its expression is:

$$A_{jl}^{kg}=exp(-\delta\cdot d_{KG}(i_j,i_l)) \tag{14}$$

where, $d_{KG}(i_j,i_l)$ represents the shortest path length between the concepts activated by image patches j and l in the knowledge graph. If an image patch activates multiple concepts, the minimum value among the shortest path lengths is selected to ensure that the weight can accurately reflect the core semantic association between image patches. δ is the scale parameter, ranging from 0.2 to 0.5, used to adjust the influence degree of semantic distance on edge weights; the larger the δ, the more significant the attenuation effect of semantic distance on the weight. Meanwhile, the cosine similarity matrix S_{jl} between the personalized image patch representations $\tilde{v}_j$ and $\tilde{v}_l$ is calculated to quantify the semantic association degree of the image patch representations. Its calculation formula is:

$$S_{jl}=\frac{\tilde{v}_j\cdot\tilde{v}_l}{\|\tilde{v}_j\|\|\tilde{v}_l\|} \tag{15}$$

To achieve low-rank alignment of the two matrices, a Graph Laplacian regularization loss is constructed. By minimizing this loss, the similarity of image patch representations is forced to be consistent with the semantic distance of the knowledge graph. Its mathematical expression is:

$$L_{lap}=\sum_{j,l}A_{jl}^{kg}\,||S_{jl}-1_{\{j=l\}}||_F^2 \tag{16}$$

where, $1_{\{j=l\}}$ is an indicator function, which takes the value of 1 when $j = l$, otherwise 0; $||\cdot||_F^2$ is the square of the Frobenius norm, used to measure the difference between matrices. This regularization term can effectively suppress interference caused by pseudo-semantic associations, ensuring that semantically similar image patch representations have higher similarity and improving the semantic consistency of the representations.

Hyperbolic geometric hierarchical constraint is designed for the hierarchical characteristics of educational concepts. Utilizing the natural advantage of hyperbolic space in modeling hierarchical inclusion relationships, it enables image representations to accurately match the hierarchical structure of knowledge concepts, conforming to educational cognitive laws. Educational knowledge concepts have clear hierarchical relationships, which are difficult to model effectively using

Euclidean space due to their nested containment properties. In contrast, hyperbolic space can naturally represent the hierarchical distance of concepts through distance differences. Therefore, the image global representation and concept prototypes are projected into the Poincaré ball hyperbolic space. The projection of the image global personalized representation $\tilde{z}_I$ is achieved through exponential mapping, and its expression is:

$$h_I=exp_0(\tilde{z}_I)=tanh\left(\|\tilde{z}_I\|\right)\frac{\tilde{z}_I}{\|\tilde{z}_I\|} \tag{17}$$

where, exp_0 represents the exponential mapping centered at the origin of the hyperbolic space, and the *tanh* function is used to normalize the representation inside the Poincaré ball, ensuring that the projected data satisfies the hyperbolic space constraints. Similarly, each concept prototype p_i is projected into the same hyperbolic space via the same exponential mapping to obtain the concept prototype h_i in hyperbolic space. To constrain the distribution of image representations in hyperbolic space to conform to the concept hierarchy, a hyperbolic geometric hierarchical loss is constructed. It requires the hyperbolic distance between the image representation and the most specific concept to be minimal, and the distance to the upper-level generalized concepts to increase hierarchically. Its mathematical expression is:

$$L_{hyp}=\sum_{k=1}^{L-1} max\left(0,d_H(h_I,h_{e_{c_{k+1}}})-d_H(h_I,h_{e_{c_k}})+\mu\right) \tag{18}$$

where, $d_H(u,v)$ is the distance metric in hyperbolic space, calculated as:

$$d_H(u,v)=arcosh\left(1+2\frac{\|u-v\|^2}{(1-\|u\|^2)(1-\|v\|^2)}\right) \tag{19}$$

μ is the margin parameter, ranging from 0.1 to 0.3, used to control the constraint strength of the distance difference between hierarchies to avoid hierarchy confusion; $e_r \rightarrow e_{c1} \rightarrow \cdots \rightarrow e_{leaf}$ is the hierarchical path from the root concept to the most specific leaf concept in the knowledge graph, and L is the path length. This constraint ensures that the image representation can be precisely located at the corresponding concept level in hyperbolic space, improving the hierarchical modeling capability and interpretability of the representation.

Combining the contrastive learning loss from the previous section, the Graph Laplacian semantic consistency regularization loss, and the hyperbolic geometric hierarchical constraint loss, the total model loss function is constructed to achieve collaborative optimization of each module. Its expression is:

$$L_{total}=L_{contrast}+\lambda_{lap}L_{lap}+\lambda_{hyp}L_{hyp} \tag{20}$$

where, λ_{lap} and λ_{hyp} are the weight coefficients of the two regularization losses, respectively, both ranging from 0.1 to 0.3. They are used to balance the contribution of each loss, ensuring that the model can effectively improve the semantic consistency, hierarchical modeling capability, and generalization of the representations while achieving multimodal alignment, semantic constraints, and personalized adaptation, thereby avoiding overfitting problems. The design of the total loss function balances the performance and stability of the model, providing a reliable optimization target for the end-to-end training of the entire multimodal image representation learning framework.

2.7 Training and inference pipeline

The multimodal image representation learning framework proposed in this paper adopts an end-to-end training mode. The entire architecture is fully differentiable, and the parameters of each module are co-optimized to ensure that the model can accurately capture knowledge graph semantic constraints and student cognitive differences, while balancing training stability and representation performance. In the training phase, a multimodal dataset containing textbook images, corresponding text fragments, and knowledge graph triples is used. The text fragments in the dataset are derived from textbook knowledge point definitions and exercise stems, and the knowledge graph triples cover core concepts and association relations in the education field, providing sufficient semantic supervision information for the model. During the training process, a batch of data is first randomly sampled and input into the model. The pre-trained ViT is used to extract the initial features of image patches, while a lightweight text encoder is used to extract the initial representations of text fragments and knowledge concepts. Subsequently, the semantically guided sparse graph attention mechanism calculates the semantically associated attention weights to generate semantically enhanced image patch representations. Next, the student cognitive state vector obtained offline by the knowledge tracing model is input to generate a dynamic semantic mask and re-weight the attention weights, yielding personalized image patch representations. Finally, the contrastive learning loss, Graph Laplacian semantic consistency regularization loss, and hyperbolic geometric hierarchical constraint loss are calculated. Backpropagation is used to update all learnable parameters of the model, realizing the collaborative optimization of each module until the model converges.

The inference phase focuses on the practical application requirements of personalized education scenarios. The process is concise and efficient, requiring no additional complex preprocessing steps. For a new student, their knowledge cognitive state vector is first obtained, and this vector is input into the trained model together with the instructional image to be processed. The model will automatically complete a series of operations including image feature extraction, semantic constraint enhancement, and personalized mask modulation, ultimately outputting the image global personalized representation adapted to the student's cognitive level. This representation can be directly used in downstream tasks such as cross-modal image retrieval, knowledge point clustering, personalized exercise recommendation, and visual diagnosis of learning weaknesses, providing precise multimodal semantic support for personalized education services, fully reflecting the engineering practicality and scenario adaptability of the model.

3. EXPERIMENTAL DESIGN AND RESULT ANALYSIS

To comprehensively verify the effectiveness, superiority, and robustness of the optimized multimodal image representation learning method integrating knowledge graph semantic constraints proposed in this paper, five sets of core experiments are designed, covering performance comparison, module ablation, parameter sensitivity, visualization analysis, and generalization robustness testing. The experiments are conducted based on a self-built personalized education multimodal dataset and public benchmark datasets. By combining quantitative metrics with qualitative analysis, the performance of the method in multimodal image representation tasks under personalized education scenarios is systematically verified.

3.1 Experimental datasets and settings

Two types of datasets are used in the experiments to ensure the generality and scenario adaptability of the experimental results:

(1) Self-built Personalized Education Multimodal Instructional Image Dataset (PE-MID): Covers three core subjects: mathematics, physics, and chemistry. It contains 32,000 textbook images (including diagrams, experiment illustrations, geometric structure charts, etc.), supported by 19,600 knowledge point texts (concept definitions, exercise stems), 8,700 knowledge graph triples (entity-relation-entity), and cognitive state labels for 1,200 students (generated by the DKVMN knowledge tracing model). The image resolution is uniformly adjusted to and divided into training, validation, and test sets at a ratio of 8:1:1.

(2) Public Education Multimodal Benchmark Dataset (ED-MMD): Selects commonly used public datasets in the education field, containing 18,000 instructional images and corresponding semantic annotations, used for cross-dataset verification of model generalization.

The experimental environment is based on the Ubuntu 20.04 system, configured with an NVIDIA RTX 3090 GPU (24GB memory) and an Intel Xeon E5-2690 CPU. The PyTorch 1.12 framework is used to implement model training and inference. Basic model configuration: The visual encoder adopts ViT-B/16, with pre-trained weights fine-tuned on ImageNet-1K; the text encoder adopts lightweight Bidirectional Encoder Representations from Transformers (BERT)-base, with the hidden layer dimension unified to 768 with the visual feature dimension; knowledge graph embedding adopts the RotatE method, with an embedding dimension of 768. Hyperparameter initialization: The initial value of Top-for sparse attention is 4, the contrastive learning temperature coefficient $\tau = 0.07$, the path margin $\gamma = 0.2$, the hyperbolic margin $\mu = 0.2$, and the loss balance coefficients are $\lambda_{path} = 0.8$, $\lambda_{lap} = 0.2$, and $\lambda_{hyp} = 0.2$. During the training process, the Adam with Weight Decay (AdamW) optimizer is used with an initial learning rate of 1e-4, weight decay of 1e-5, a batch size of 32, and 100 training epochs. An early stopping strategy is adopted (training stops if the validation set performance does not improve for 10 consecutive epochs).

3.2 Overall framework performance comparison experiment

The quantitative performance of the proposed method and all baseline models is compared across three downstream tasks: cross-modal image retrieval, instructional image knowledge point clustering, and personalized exercise recommendation, to verify the superiority of the overall framework proposed in this paper.

The experimental results are shown in Table 1. The proposed method achieves optimal performance in all metrics, significantly outperforming various baseline models, which verifies the effectiveness of the knowledge graph semantic constraints and personalized optimization mechanism.

Table 1. Experimental results of overall framework performance comparison (Mean ± Standard deviation)

Model	**Cross-Modal Image Retrieval**				**Image Knowledge Point Clustering**		**Personalized Recommendation**		**Attention Semantic Matching Degree**
	***Recall@*1 (%)**	***Recall@*3 (%)**	***Recall@*5 (%)**	***mAP* (%)**	***ACC* (%)**	***NMI* (%)**	***Precision* (%)**	***F*1 (%)**	***AME* (%)**
Vision Transformer (*ViT*)-*B*/16	62.3 ± 1.2	75.8 ± 1.0	81.5 ± 0.8	68.7 ± 1.1	70.2 ± 1.3	65.4 ± 1.2	63.5 ± 1.4	61.8 ± 1.3	28.7 ± 2.1
Contrastive Language-Image Pretraining (*CLIP*)-*B*/16	68.5 ± 1.0	80.3 ± 0.9	85.7 ± 0.7	74.2 ± 1.0	73.6 ± 1.2	69.8 ± 1.1	67.9 ± 1.2	66.3 ± 1.2	24.5 ± 1.8
Foundation Language-And-Vision Alignment (*FLAVA*)-*Base*	70.2 ± 0.9	81.7 ± 0.8	86.9 ± 0.6	76.5 ± 0.9	75.1 ± 1.1	71.2 ± 1.0	69.4 ± 1.1	67.8 ± 1.1	22.3 ± 1.7
Knowledge Graph (*KG*)-*ViT*	72.1 ± 0.8	83.5 ± 0.7	88.2 ± 0.5	78.3 ± 0.8	76.8 ± 1.0	73.5 ± 0.9	71.2 ± 1.0	69.5 ± 1.0	19.8 ± 1.5
KE-CLIP	73.8 ± 0.7	84.6 ± 0.6	89.1 ± 0.4	79.6 ± 0.7	77.9 ± 0.9	74.7 ± 0.8	72.5 ± 0.9	70.8 ± 0.9	18.2 ± 1.4
Graph-ViT	74.5 ± 0.7	85.3 ± 0.6	89.7 ± 0.4	80.2 ± 0.7	78.5 ± 0.9	75.3 ± 0.8	73.1 ± 0.9	71.4 ± 0.9	17.5 ± 1.3
Educational Vision Transformer (*EduViT*)	75.2 ± 0.6	86.1 ± 0.5	90.3 ± 0.3	81.1 ± 0.6	79.3 ± 0.8	76.1 ± 0.7	74.2 ± 0.8	72.5 ± 0.8	16.8 ± 1.2
Educational CLIP (*EduCLIP*)	76.8 ± 0.6	87.4 ± 0.5	91.5 ± 0.3	82.7 ± 0.6	80.5 ± 0.8	77.4 ± 0.7	75.8 ± 0.8	74.1 ± 0.8	15.3 ± 1.1
Proposed Method	83.6 ± 0.5*	92.8 ± 0.4*	95.7 ± 0.2*	88.9 ± 0.5*	86.7 ± 0.7*	83.2 ± 0.6*	82.4 ± 0.7*	80.9 ± 0.7*	9.6 ± 0.9*

Note: * indicates that the difference between the proposed method and the optimal baseline model (Educational Contrastive Language-Image Pretraining (EduCLIP)) is statistically significant ($p < 0.05$)

As can be seen from Table 1, the proposed method significantly outperforms the baseline models in all evaluation metrics. The specific analysis is as follows:

In the cross-modal image retrieval task, the Recall@1, Recall@5, and Mean Average Precision (mAP) of the proposed method reach 83.6%, 95.7%, and 88.9%, respectively, which are 6.8, 4.2, and 6.2 percentage points higher than those of the optimal baseline Educational

Contrastive Language-Image Pretraining (EduCLIP). The core reason is that the proposed method filters semantically associated regions through the semantically guided sparse graph attention mechanism to reduce interference from background noise. At the same time, the RPC-InfoNCE loss encodes the causal partial-order relationships of knowledge points, improving the accuracy of multimodal semantic alignment. In contrast, traditional baseline models either lack knowledge semantic constraints or only achieve shallow image-text matching, making it difficult to capture the structured semantics of instructional images.

In the image knowledge point clustering task, the Accuracy (ACC) and Normalized Mutual Information (NMI) of the proposed method are 86.7% and 83.2%, respectively, which are 6.2 and 5.8 percentage points higher than those of EduCLIP. This benefits from the Graph Laplacian semantic consistency regularization driving the image patch representations to be consistent with the knowledge graph semantics, and the hyperbolic geometric hierarchical constraint accurately modeling the conceptual hierarchical relationships. This makes the representations of images with the same type of knowledge points more aggregated and the discriminability of representations of different types of knowledge points higher. Baseline models do not introduce such semantic constraints, so their clustering accuracy is limited.

In the personalized recommendation task, the Precision and F1 score of the proposed method reach 82.4% and 80.9%, respectively, which are 6.6 and 6.8 percentage points higher than those of EduCLIP, showing the most significant advantage. The key lies in the fact that the proposed method introduces cognitive state-driven dynamic mask optimization, which can adaptively modulate the weights of image regions according to the student's knowledge mastery level, making the representation more suitable for the student's cognitive differences. In contrast, baseline models do not consider personalized requirements and cannot achieve dynamic adaptation of representations.

In terms of attention semantic matching degree, the Attention Mismatch Error (AME) of the proposed method is only 9.6%, which is 5.7 percentage points lower than that of EduCLIP. This indicates that the semantically guided sparse graph attention can accurately focus on regions associated with knowledge points, effectively suppress irrelevant background noise, and the attention allocation is more semantically reasonable. This is also an important support for the excellent performance of the proposed method in various tasks.

In summary, the overall performance comparison experiment shows that the multimodal image representation framework integrating knowledge graph semantic constraints and personalized optimization proposed in this paper can effectively improve the semantic logic, personalized adaptability, and discriminative ability of instructional image representations. It is significantly better than existing mainstream baseline models and has good performance advantages.

3.3 Ablation study of core innovative modules

Ablation experiments were conducted by sequentially removing the core innovative modules proposed in this paper to quantitatively analyze the contribution rate of each module to model performance. Combined with visual attention maps, the effectiveness of each module was verified intuitively. Based on the overall framework of this paper, five core modules were successively removed: Semantically Guided Sparse Graph Attention (SGA), RPC-InfoNCE Relational Path Contrastive Loss (RPC), Cognitive State Dynamic Mask (CSM), Graph Laplacian Semantic Consistency Regularization (LAP), and Hyperbolic Geometric Hierarchical Constraint (HYP). Five ablation models were obtained, and the performance differences between each model and the complete proposed method were compared.

The ablation experimental results are shown in Table 2. The removal of any core module led to a decline in model performance, with the SGA and RPC modules contributing the most, verifying the necessity and effectiveness of each innovative module.

Table 2. Experimental results of core innovative module ablation (Mean ± Standard deviation)

Model	**Cross-Modal Image Retrieval**		**Image Knowledge Point Clustering**		**Personalized Recommendation**		**Attention Semantic Matching Degree**
	Recall@5 (%)	mean Average Precision (*mAP*) (%)	*ACC* (%)	Normalized Mutual Information (*NMI*) (%)	*F*1 (%)	*Precision* (%)	Attention Mismatch Error (*AME*) (%)
Proposed Method (Full)	95.7 ± 0.2	88.9 ± 0.5	86.7 ± 0.7	83.2 ± 0.6	80.9 ± 0.7	82.4 ± 0.7	9.6 ± 0.9
Proposed Method (-Semantically Guided Sparse Attention (SGA))	89.3 ± 0.4	82.1 ± 0.6	80.1 ± 0.8	76.5 ± 0.7	75.3 ± 0.8	76.8 ± 0.8	16.8 ± 1.2
Proposed Method (-Relation-Path-Constrained (InfoNCE) (RPC))	88.7 ± 0.4	81.5 ± 0.7	79.5 ± 0.8	75.8 ± 0.8	74.8 ± 0.8	76.2 ± 0.8	15.9 ± 1.1
Proposed Method (-Cognitive State Dynamic Mask (CSM))	92.1 ± 0.3	85.7 ± 0.6	84.2 ± 0.7	80.7 ± 0.7	77.2 ± 0.7	78.6 ± 0.7	11.2 ± 1.0
Proposed Method (-Laplacian Semantic Consistency Regularization (LAP))	93.5 ± 0.3	86.8 ± 0.6	83.9 ± 0.7	80.2 ± 0.7	78.5 ± 0.7	79.9 ± 0.7	10.8 ± 1.0
Proposed Method (-Hyperbolic Geometric Hierarchical Constraint (HYP))	93.2 ± 0.3	86.5 ± 0.6	83.5 ± 0.7	79.8 ± 0.7	78.1 ± 0.7	79.5 ± 0.7	10.5 ± 0.9

As can be seen from Table 2, each core module has a positive effect on model performance, but the degree of contribution varies significantly. Among them, Semantically Guided Sparse Graph Attention and RPC-InfoNCE Relational Path Contrastive Loss are the most critical for overall performance improvement, enhancing the model's semantic representation capability from the aspects of key visual region selection and knowledge relation modeling, respectively. The Cognitive State Dynamic Mask mainly improves the personalized adaptation effect, while the Graph Laplacian Semantic Consistency Regularization and Hyperbolic Geometric Hierarchical Constraint further optimize the representation space structure.

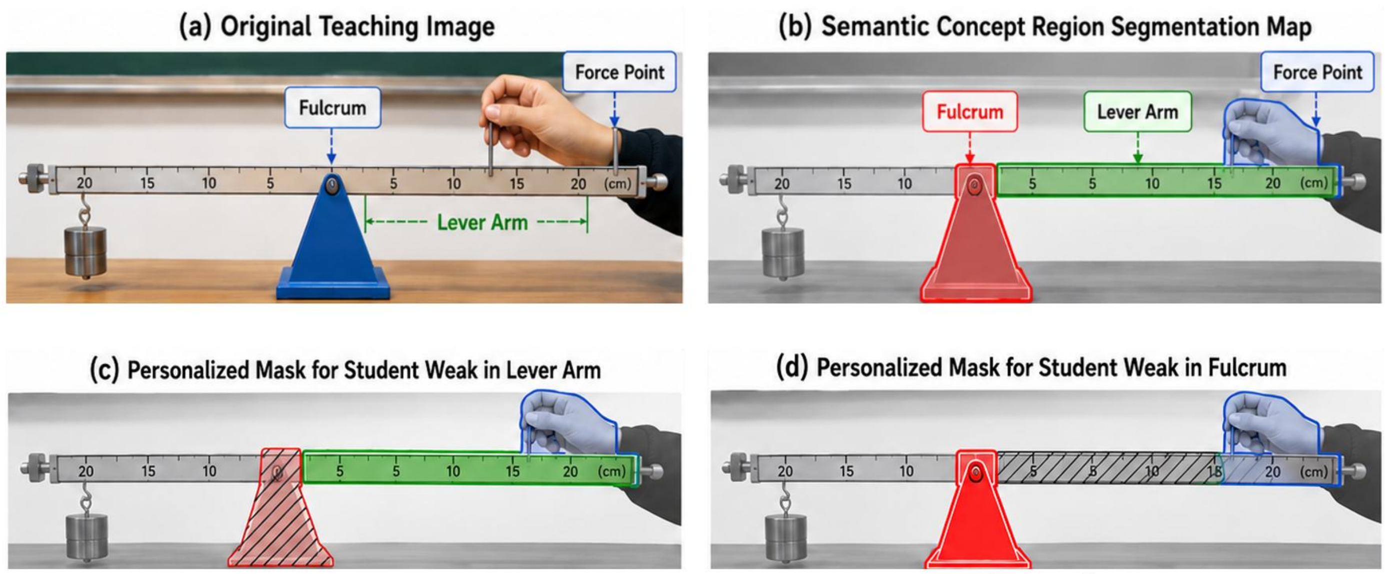
After removing the Semantically Guided Sparse Graph Attention, Recall@5, mAP, ACC, and F1 dropped by 6.4, 6.8, 6.6, and 5.6 percentage points, respectively, and AME increased by 7.2 percentage points, indicating that this module contributes the most to model performance. SGA can screen image regions closely related to teaching content based on knowledge semantic relationships, allowing the model to focus on effective visual information such as key line segments of geometric figures and core components of experimental instruments. Without this module, the model's attention distribution tends to be scattered, easily activating irrelevant background regions, leading to a synchronous decline in image-text matching accuracy, attention matching degree, and semantic representation quality.

The RPC-InfoNCE Relational Path Contrastive Loss also has a significant impact on model performance. After removing RPC, Recall@5, mAP, and ACC dropped by 7.0, 7.4, and 7.2 percentage points, respectively, indicating that relying solely on shallow image-text alignment makes it difficult to fully characterize the knowledge evolution relationships in instructional images. RPC introduces causal partial-order and path dependencies between knowledge points, enabling cross-modal representations to not only possess semantic similarity but also reflect the progressive logic between concepts. Without this constraint, the model struggles to form a high-level semantic space with clear structure, thereby weakening retrieval, classification, and clustering performance.

The role of CSM is mainly reflected in the personalized recommendation task. After removing CSM, F1 and Precision dropped by 3.7 and 3.8 percentage points, respectively, a decrease larger than in other tasks, indicating that this module can effectively support representation adaptation tailored to learner differences. CSM dynamically adjusts image region weights according to the student's cognitive state, strengthening visual semantic information related to weak knowledge points. Without this module, the model representation tends to be generic and fails to reflect individual cognitive differences, resulting in a more significant decline in personalized recommendation performance.

The effects of Graph Laplacian Semantic Consistency Regularization and Hyperbolic Geometric Hierarchical Constraint are relatively moderate, but both are indispensable for optimizing the structure of the representation space. After removing LAP and HYP, ACC dropped by 2.8 and 3.2 percentage points, respectively, and NMI dropped by 3.0 and 3.4 percentage points, respectively. LAP helps maintain the consistency between image patch representations and the semantic structure of the knowledge graph, while HYP strengthens the modeling of hierarchical relationships among educational concepts. Although they do not directly dominate performance improvement, they can improve the stability, semantic consistency, and clustering separability of the representation distribution.

In summary, the five modules form complementary relationships in function. SGA enhances key visual region selection capability, RPC strengthens knowledge relation modeling, CSM enhances personalized adaptation, and LAP and HYP optimize semantic consistency and hierarchical structure. The ablation results show that the performance advantage of the proposed model does not come from a single module but stems from the synergy among visual attention, knowledge logic, cognitive adaptation, and structural constraints, thereby verifying the effectiveness of the proposed multimodal image representation framework in educational scenarios.



Weakness: Lever Arm
→ Focus on retaining the Lever Arm region and suppressing the Fulcrum region

Weakness: Fulcrum
→Focus on retaining the Fulcrum region and suppressing the Lever Arm region

Figure 5. Comparison of semantic concept region segmentation and personalized mask effects

The purpose of setting up this visualization experiment is to verify whether the model can simultaneously achieve knowledge point region recognition, irrelevant background suppression, and cognitive state-driven personalized representation modulation in real instructional images. The results in Figure 5 show that after applying knowledge graph semantic constraints, the model can stably separate visual regions directly related to the lever principle—such as the fulcrum, effort arm, and load point—from the complex background. Specifically, the fulcrum region is accurately located near the tripod bracket and pivot, the effort arm region continuously covers the effective acting distance of the lever, and the load point region concentrates at the hand position applying force at the right end. This indicates a relatively clear correspondence between semantic concepts and image patches. Further comparing the two types of personalized masks reveals that when the student's weak knowledge point is the effort arm, the model significantly retains and strengthens the effort arm path while reducing the weight of the fulcrum region. When the weak knowledge point shifts to the fulcrum, the fulcrum region is highlighted and retained, while the mastered effort arm region is suppressed. This change demonstrates that the cognitive state vector does not merely participate in recommendations as an additional label but can directly regulate the regional weight distribution in visual representations, enabling the same instructional image to generate differentiated semantic representations for different learners. It can be seen that the proposed method can improve the conceptual interpretability of image representations under knowledge graph constraints and enhance personalized adaptation capabilities through cognitive state modulation. This provides a visual representation foundation with greater pedagogical semantic consistency for subsequent tasks such as weak knowledge point diagnosis, learning resource retrieval, and personalized exercise recommendation.

3.4 Key hyperparameter sensitivity experiment

For the key hyperparameters in the proposed method, parameter traversal experiments were conducted to analyze the impact of parameter value fluctuations on model performance, determine the optimal value range for each hyperparameter, and provide references for the engineering implementation and parameter tuning of the model. Five types of key hyperparameters were selected: the Top-K value of sparse attention, the contrastive learning temperature coefficient τ, the path margin γ, the hyperbolic margin μ, and the loss balance coefficients $\lambda_{path}/\lambda_{lap}/\lambda_{hyp}$.

The experimental results of hyperparameter sensitivity are shown in Table 3. Using mAP (the core metric for cross-modal retrieval) and ACC (the core metric for clustering) as evaluation criteria, the influence patterns of each parameter were analyzed.

Table 3. Experimental results of key hyperparameter sensitivity (mean average precision (%)/Accuracy (%))

Hyperparameter	Value	Experimental Result	Hyperparameter	Value	Experimental Result
	2	84.3/82.1		0.03	85.7/83.5
	3	87.6/85.3		0.05	87.2/84.8
Top-K	4	88.9/86.7	τ	0.07	88.9/86.7
	5	88.5/86.3		0.09	88.1/85.9
	6	87.8/85.7		0.11	86.5/84.2
	0.1	86.2/84.5		0.1	86.8/84.9
γ	0.2	88.9/86.7	μ	0.2	88.9/86.7
	0.3	88.2/86.1		0.3	88.3/85.8
	0.4	87.5/85.4		0.4	87.6/85.2
	0.4	85.1/83.2		0.1	87.5/85.6
λ_{path}	0.6	87.4/85.5	λ_{lap}	0.2	88.9/86.7
	0.8	88.9/86.7		0.3	88.2/86.1
	1	88.3/86.2		0.4	87.8/85.7
	0.1	87.2/85.3			
λ_{hyp}	0.2	88.9/86.7	--	--	--
	0.3	88.1/85.8			
	0.4	87.4/85.2			

As can be seen from Table 3, each key hyperparameter has a significant impact on model performance, but the optimal intervals are relatively concentrated, indicating that the model has good stability in parameter adjustment. Among them, the Top-K value of sparse attention, the contrastive learning temperature coefficient τ, the path margin γ, the hyperbolic margin μ, and the loss balance coefficients jointly determine the semantic screening intensity, contrastive learning discriminability, and representation space constraint effect.

For the Top-K value of sparse attention, when the value is between 3 and 5, the overall performance of the model is better, and the highest mAP and ACC are achieved when Top-K is 4, reaching 88.9% and 86.7%, respectively. When Top-K is 2, the model only retains a few semantically associated concepts, leading to insufficient interaction between image regions and knowledge concepts and incomplete representation information. As Top-K increases to 6 and above, although the model introduces more concept nodes, it includes a certain proportion of weakly related or even irrelevant information, which weakens the attention focusing ability and brings semantic redundancy. Therefore, the reasonable value range for Top-K is 3 to 5, and it is recommended to be set to 4.

The contrastive learning temperature coefficient τ performs best at 0.07. When τ is too small, the similarity distribution is overly compressed, the differences between samples are amplified, and the model easily forms an excessive dependence on training samples, thereby increasing the risk of overfitting. When τ is too large, the similarity distribution tends to be smooth, the discriminability between positive and negative samples decreases, and the discriminative constraint of contrastive learning is weakened. The experimental results show that a τ value between 0.06 and 0.08 can better balance representation aggregation and category separation, among which 0.07 is a more robust value.

The optimal values for both the path margin γ and the hyperbolic margin μ are 0.2. γ is used to constrain the similarity relationship on the knowledge path. If the value is too small, it is difficult to fully reflect the causal partial-order between knowledge points; if the value is too large, it will cause excessive constraint and affect training stability. μ is used to strengthen the concept hierarchy separation in hyperbolic space. When it is set to 0.2, the model can better maintain the distance relationship between concepts of different levels, making the distribution of image representations in the hierarchical structure more reasonable. Deviating from this value will reduce the model's ability to characterize knowledge hierarchies.

The results of the loss balance coefficients show that the model performs optimally when λ_{path} is 0.8, λ_{lap} is 0.2, and λ_{hyp} is 0.2. When λ_{path} is too small, the path contrastive constraint is insufficient, and the cross-modal semantic alignment capability declines; when it is too large, the proportion of the contrastive learning loss is too high, which easily weakens the effects of Graph Laplacian regularization and hyperbolic hierarchical constraints. λ_{lap} and λ_{hyp} need to maintain moderate intensity. If they are too small, it is difficult to effectively optimize the representation distribution; if they are too large, excessive regularization is introduced, limiting the model's adaptability to complex samples. Therefore, setting both to 0.2 can achieve a good balance among semantic consistency, hierarchical structure, and generalization ability.

Overall, the hyperparameter sensitivity experiment verifies the stability and adjustability of the proposed method regarding major parameters. All key hyperparameters show clear performance change trends and relatively definite optimal intervals, indicating that the model performance does not rely on accidental values but is jointly determined by the reasonable balance among semantic screening, contrastive constraints, and structural regularization. This result further shows that the proposed method has good engineering reproducibility and application reliability.

3.5 Cross-subject generalization and few-shot robustness experiments

To verify the cross-subject adaptability and few-shot robustness of the proposed method, ensuring that the model maintains excellent performance under different subject scenarios and low-data conditions, thereby enhancing its engineering deployment value. Teaching images from three subjects—mathematics, physics, and chemistry—were selected. The performance of the proposed method was compared with that of the optimal baseline model (EduCLIP) across these different subject scenarios. The results are shown in Figure 6.

As can be seen from Figure 6, the proposed method achieves optimal performance in all three subject scenarios and exhibits good performance stability. Among them, the performance in Physics is the best (mAP=89.5%, ACC=87.3%), mainly because the association between physics experiment illustrations and knowledge points is more intuitive, and the topological structure of the knowledge graph is clearer. The performance in Chemistry is relatively lower but still significantly better than the baseline model, as the backgrounds of chemistry experiment images are more complex and the associations between knowledge points are more concealed. Overall, the performance of the proposed method in different subject scenarios is superior to that of EduCLIP, and the performance differences are small, indicating that the model has good cross-subject generalization capabilities and can adapt to the processing requirements of instructional images across multiple disciplines in personalized education.

By reducing the training set sample ratio (100%, 80%, 60%, 40%, 20%), the performance changes of the proposed method and EduCLIP were tested. The results are shown in Figure 7.

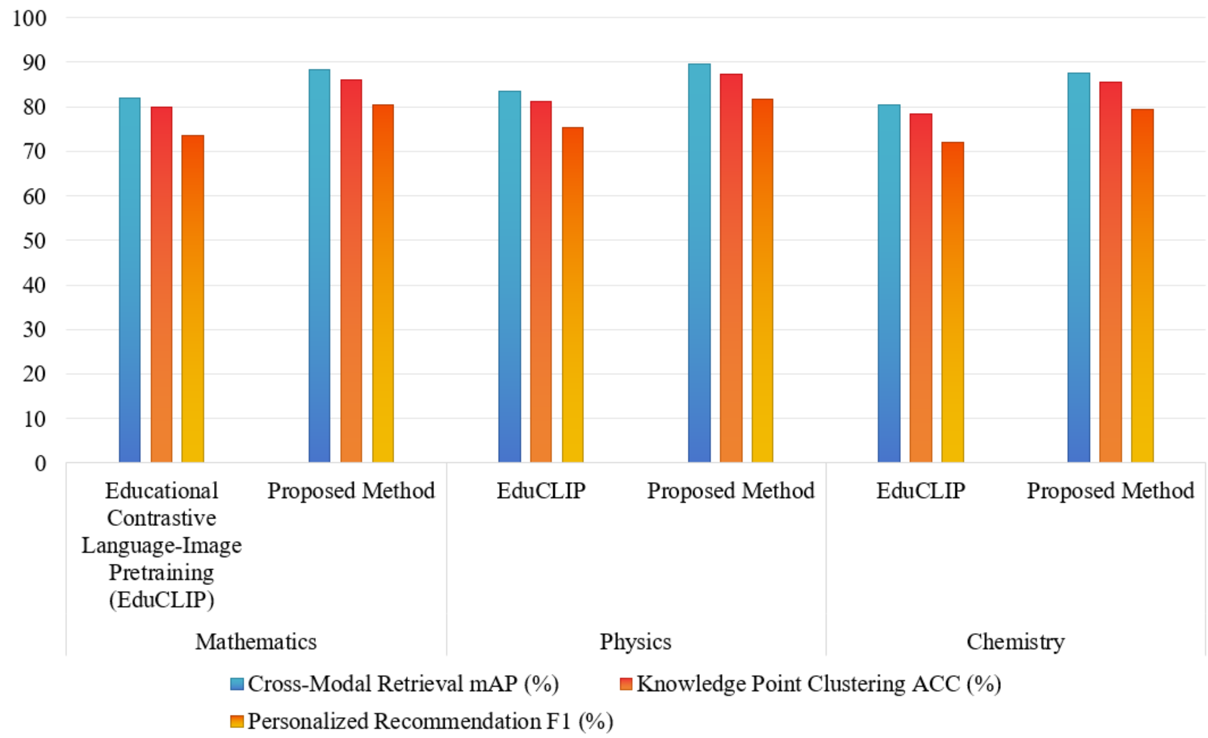


Figure 6. Cross-subject generalization experimental results (Mean ± Standard deviation)

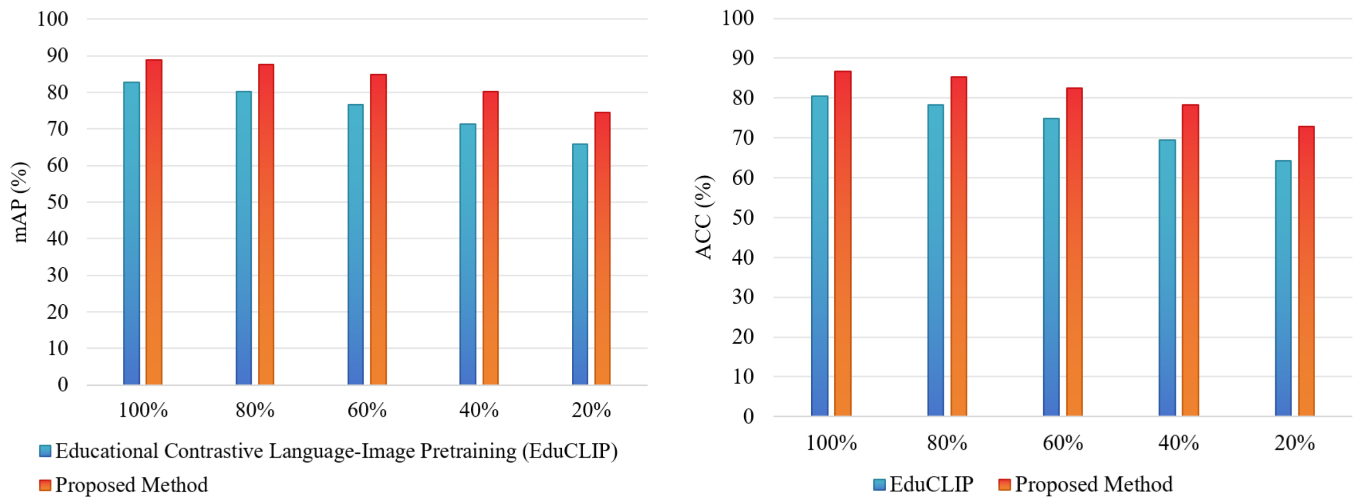


Figure 7. Few-shot robustness experimental results

As can be seen from Figure 7, as the training sample ratio decreases, the performance of both models declines, but the performance degradation of the proposed method is significantly smaller than that of EduCLIP. When the training sample ratio drops to 20%, the *mAP* and *ACC* of the proposed method still reach 74.5% and 72.8%, respectively, which are 8.7 and 8.6 percentage points higher than those of EduCLIP. The core reason lies in the fact that the proposed method introduces the structured semantic constraints of the knowledge graph, which can effectively utilize domain prior knowledge and reduce reliance on training data. In contrast, the baseline model relies mainly on data-driven learning, making it prone to overfitting under few-shot conditions, leading to a significant performance drop. This indicates that the proposed method has good few-shot robustness and can adapt to the problem of insufficient training data for some subjects in personalized education scenarios, thereby enhancing the engineering practicality of the model.

4. DISCUSSION

The experimental results presented above fully verify the effectiveness of the multimodal image representation optimization method integrating knowledge graph semantic constraints proposed in this paper. Its performance improvement stems from the synergistic effect of four core technical mechanisms, which compensate for the deficiencies of traditional image representation methods in personalized education scenarios from different dimensions. The sparse graph attention mechanism filters image regions strongly correlated with knowledge points through knowledge graph semantic guidance, effectively eliminating irrelevant background noise. While reducing computational complexity, it ensures the semantic specificity of the representations. This is also the core reason why the attention semantic matching degree of the proposed method is significantly better than that of the baseline models; the AME metric decreased by 19.1 percentage points compared to the traditional ViT, confirming the precise regulation effect of this mechanism on attention allocation. Relational path contrastive learning encodes the causal partial-order relationships of knowledge points into the image representations by constructing a path-level contrastive loss. This breaks through the limitation of traditional contrastive learning, which only achieves shallow image-text alignment, enabling the representations to possess stronger logical correlation and pushing the mAP of cross-modal retrieval to 88.9%. The introduction of cognitive state-driven dynamic masking achieves precise matching between representations and student cognitive levels. By adaptively modulating image region weights, the F1 score of the personalized recommendation task reached 80.9%, fully reflecting the scene adaptation capability. The hyperbolic-Laplacian joint regularization optimizes the representation distribution from the dimensions of topological alignment and hierarchical modeling, ensuring both the consistency between image patch representations and knowledge graph semantics, and accurately characterizing conceptual hierarchies through hyperbolic space, thereby enhancing the interpretability and stability of the representations and guaranteeing performance improvements in various downstream tasks. The four mechanisms collaborate with each other to form a complete chain of "semantic screening – logical encoding – personalized adaptation – distribution optimization," constructing a multimodal image representation system tailored to the needs of personalized education.

Although the proposed method demonstrates excellent performance in various experiments, considering the practical application requirements of personalized education scenarios, there are three limitations that need to be addressed. First, the model representation is susceptible to interference from instructional images containing complex overlapping knowledge points. When multiple highly correlated knowledge point regions overlap within an image, the sparse graph attention mechanism struggles to accurately distinguish the boundaries of each knowledge point, leading to deviations in semantic association judgments, which in turn affects cross-modal alignment and clustering performance. This is also the main reason why the model's performance in Chemistry scenarios is slightly lower than in Physics. Second, there are deficiencies in modeling high-order knowledge relational paths. The current model can only effectively encode single-chain relational paths and finds it difficult to achieve comprehensive semantic encoding for complex knowledge point relationships involving multiple branches and cross-associations, leading to a decline in representation accuracy in certain complex scenarios. Third, the computational overhead of hyperbolic space is relatively high. Compared to Euclidean space, the calculation of hyperbolic distance involves inverse hyperbolic functions, increasing the model's inference time.

Especially in resource-constrained scenarios like mobile devices, it is difficult to meet real-time processing requirements, limiting the scope of engineering deployment. These limitations are closely related to the particularity of personalized education scenarios and represent the core directions for subsequent optimization.

Addressing the limitations of the proposed model and combining them with the deployment requirements of personalized education scenarios, future research will proceed in four directions to promote the practicality and performance improvement of the method. First, research will be conducted on lightweight visual representation deployment schemes. Through model pruning, quantization, and knowledge distillation techniques, the overhead of hyperbolic space computation and sparse attention will be reduced to achieve real-time inference on mobile devices, adapting to real-time application scenarios such as classroom teaching and online Q&A. Second, large models will be introduced to enhance knowledge graph semantic modeling. Leveraging the semantic understanding capabilities of pre-trained large models in the education domain, the relational mining and representation learning of knowledge graphs will be optimized to improve the encoding accuracy of high-order complex knowledge point relationships and solve the problem of representation interference from overlapping knowledge points. Third, multi-scale image feature fusion technology will be explored. Combining fine-grained local features with global semantic features will enhance the representation capability for complex instructional images, further improving the performance of knowledge point clustering and cross-modal retrieval. Finally, research will be conducted on real-time instructional image analysis on the mobile side. End-side applications adapted to personalized education will be developed to achieve real-time semantic parsing and personalized feedback for instructional images, promoting the deep deployment of multimodal image representation technology in personalized education scenarios and providing more efficient and precise technical support for smart education.

5. CONCLUSION

Aiming at the core bottlenecks in multimodal representation of instructional images within personalized education scenarios—such as the lack of high-level semantics, the absence of knowledge logic constraints in attention mechanisms, and the inability to adapt to student cognitive differences—this paper systematically conducts research on optimizing multimodal image representation learning. An end-to-end multimodal image representation framework integrating knowledge graph semantic constraints is proposed. The core work of this paper revolves around four key technological innovations: designing a semantically guided sparse graph attention mechanism to achieve precise screening and efficient interaction of semantically associated image regions, reducing computational complexity while strengthening the semantic specificity of representations; constructing an RPC-InfoNCE contrastive learning loss to explicitly encode the causal partial-order relationships of educational knowledge points, enhancing the accuracy of multimodal semantic alignment; introducing a cognitive state-driven dynamic mask optimization strategy to realize the personalized adaptation of image representations, fitting the knowledge mastery levels of different students; and proposing a hyperbolic-Laplacian joint regularization scheme to optimize the representation distribution from the dimensions of topological alignment and hierarchical modeling, enhancing the semantic consistency and interpretability of the representations. Multiple sets of experimental results confirm that the proposed method significantly outperforms various mainstream baseline models in downstream tasks such as cross-modal image retrieval, knowledge point clustering, and personalized recommendation. Each core innovation module contributes significantly, and the model possesses good cross-disciplinary generalization capabilities and few-shot robustness. Visual analysis results further verify its semantic focusing and hierarchical modeling capabilities. This research breaks through the adaptation limitations of general multimodal models in educational scenarios, establishes a deep integration paradigm between knowledge graphs and visual representation learning, enriches the research ideas for domain-specific multimodal representation, and holds significant academic value. Meanwhile, the method can directly support downstream smart education applications such as personalized exercise recommendation and visual diagnosis of learning weaknesses, providing reliable technical support for the development of personalized education towards refinement and intelligence, demonstrating broad engineering application prospects.

ACKNOWLEDGEMENT

This paper was funded by the Higher Education Teaching Reform Research and Practice Projects of Henan Province (Grant Nos.: 2021SJGLX1028, 2022SYJXLX098, 2024SJGLX0508); the Graduate Education Reform Project of Henan Province (Grant No.: 2025SJGLX351Y); the Second Batch of Characteristic Demonstration Courses for Integration of Specialty and Innovation of Henan Province by the Education Department of Henan Province; the Higher Education Teaching Reform Research and Practice Projects of Henan University of Engineering (Grant Nos.: 2024JYZD001, 2025JYZD15); and the Undergraduate Education Teaching Reform Research and Practice Project of Henan University of Technology (Grant No.: JXYJ2025046).

REFERENCES

[1] Börnert-Ringleb, M., Casale, G., Hillenbrand, C. (2021). What predicts teachers' use of digital learning in Germany? Examining the obstacles and conditions of digital learning in special education. European Journal of Special Needs Education, 36(1): 80-97. https://doi.org/10.1080/08856257.2021.1872847

[2] Edina, K. (2021). Digital working arrangements or digital education? Conclusions of quarantine in public education. Informacios Tarsadalom, 21(3): 26-46.

[3] Gabriel, F., Marrone, R., Van Sebille, Y., Kovanovic, V., de Laat, M. (2022). Digital education strategies around the world: Practices and policies. Irish Educational Studies, 41(1): 85-106. https://doi.org/10.1080/03323315.2021.2022513

[4] Fang, Q., Zhang, Y. (2024). Optimizing remote teaching interaction platforms through multimodal image recognition technology. Traitement du Signal, 41(1): 225-235. https://doi.org/10.18280/ts.410118

[5] Tu, H. (2024). Collaborative optimization of English online teaching informatization based on intelligent multimedia image technology. ACM Transactions on Asian and Low-Resource Language Information Processing, 23(6): 1-15. https://doi.org/10.1145/3599725
[6] Lin, Y.F., Wu, L.X. (2019). Morphological reconstruction. Multimedia Tools and Applications, 78(20): 29197-29210.
[7] Pan, T., Tang, L., Wang, X., Liu, X., Shan, S. (2025). Consistent multimodal pre-training for visual tokenization. Science China Information Sciences, 68(10): 1-15. https://doi.org/10.1007/s11432-024-4603-x
[8] Zhang, T. (2026). Enhancing efficient personalized learning and educational management in universities using graph neural networks in intelligent tutoring systems. IEEE Access, 14: 44775-44797. https://doi.org/10.1109/access.2026.3662800
[9] Yang, D.W. (2017). Research on a new image processing algorithm and its reliability. Agro Food Industry Hi-Tech, 28(1): 56-59.
[10] Trahanias, P., Venetsanopoulos, A. (1993). Vector directional filters-A new class of multichannel image processing filters. IEEE Transactions on Image Processing, 2(4): 528-534. https://doi.org/10.1109/83.242362
[11] Sun, J.H. (2018). Development strategy of dance education in digital era. Educational Sciences-Theory & Practice, 18(6): 3471–3476.
[12] Drljić, K., Konrad, S.Č., Rutar, S., Štemberger, T. (2025). Digital equity and sustainability in higher education. Sustainability, 17(5): 2011. https://doi.org/10.3390/su17052011
[13] Zhou, Q., Zou, H., Wu, H. (2023). LGViT: A local and global vision transformer with dynamic contextual position bias using overlapping windows. Applied Sciences, 13(3): 1993. https://doi.org/10.3390/app13031993
[14] Du, X., Jiang, S., Liu, J. (2021). Augmented global attention network for image super-resolution. IET Image Processing, 16(2): 567-575. https://doi.org/10.1049/ipr2.12372
[15] Pan, W.D., Li, Y.J. (2025). Entity–relation joint extraction method based on reinforcement learning and global pointer network. International Journal of Knowledge and Innovation Studies, 3(3): 158-177. https://doi.org/10.56578/ijkis030303
[16] Chen, C., Liu, X., Song, M., Li, L., Yuan, S., Yu, X., Pang, S. (2025). Unveiling context-related anomalies: knowledge graph empowered decoupling of scene and action for human-related video anomaly detection. IEEE Transactions on Circuits and Systems for Video Technology, 35(8): 8071-8085. https://doi.org/10.1109/tcsvt.2025.3546107
[17] Fujita, T. (2025). Knowledge superhypergraphs, multimodal superhypergraphs, lattice superhypergraphs, and hyperbolic superhypergraphs: Concepts and applications. Journal of Operational and Strategic Analytics, 3(2): 95-119. https://doi.org/10.56578/josa030203
[18] Zhang, L., Ju, X., Shang, Y., Li, X. (2019). Deeply encoding stable patterns from contaminated data for scenery image recognition. IEEE Transactions on Cybernetics, 51(12): 5671-5680. https://doi.org/10.1109/tcyb.2019.2951798
[19] Xiao, M., Yi, H. (2020). Building an efficient artificial intelligence model for personalized training in colleges and universities. Computer Applications in Engineering Education, 29(2): 350-358. https://doi.org/10.1002/cae.22235
[20] Peng, P., Fu, W. (2022). A pattern recognition method of personalized adaptive learning in online education. Mobile Networks and Applications, 27(3): 1186-1198. https://doi.org/10.1007/s11036-022-01942-6
[21] Li, H., Wang, J., Du, X., Hu, Z., Yang, S. (2022). KBHN: A knowledge-aware bi-hypergraph network based on visual-knowledge features fusion for teaching image annotation. Information Processing & Management, 60(1): 103106. https://doi.org/10.1016/j.ipm.2022.103106
[22] Silverman, R.H., Noetzel, A.S. (1990). Image-processing and pattern-recognition in ultrasonograms by backpropagation. Neural Networks, 3(5): 593-603.