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

Traitement du Signal
Vol. 42, No. 5, October, 2025, pp. 2585-2596
Journal homepage: http://iieta.org/journals/ts

A Collaborative Learning Framework and Optimization Approach for Visual Tasks and Knowledge Graphs in Intelligent Education

Caiying Lan[1], HaoxinYang[2*]

[1] School of Marxism, Inner Mongolia University of Technology, Hohhot 010051, China
[2] School of Physical Education, Inner Mongolia Normal University, Hohhot 010022, China

Corresponding Author Email: 15949474954@163.com

https://doi.org/10.18280/ts.420512

Received: 11 January 2025
Revised: 3 August 2025
Accepted: 18 August 2025
Available online: 31 October 2025

Keywords:
intelligent education, visual tasks, knowledge graph, multimodal fusion, contrastive learning, cross-modal attention

ABSTRACT

In the context of deep integration between information technology and education, intelligent education scenarios generate massive amounts of multimodal data. While knowledge graphs can organize and associate such data, the synergy between visual tasks and knowledge graphs remains insufficient, limiting the full exploitation of multimodal knowledge and constraining the accuracy and intelligence of visual tasks. Existing studies on combining multimodal knowledge processing with visual tasks in intelligent education exhibit notable shortcomings: they fail to effectively resolve semantic inconsistencies across modalities, rely on inflexible methods to extract cross-modal associations, and lack frameworks with strong generality and scalability. To address these issues, this study undertakes three major research efforts: (1) employing contrastive learning to reduce semantic inconsistencies between modalities and enhance the discriminative ability of multimodal embeddings for the same entity, thereby achieving feature enhancement; (2) designing a cross-modal attention module to extract complementary information across modalities and optimize textual features with image features; and (3) developing a general and scalable collaborative learning framework that integrates multimodal prediction results through joint decision-making to improve link prediction accuracy. The innovations of this work lie in: effectively alleviating cross-modal semantic inconsistencies via contrastive learning to improve feature representation accuracy; dynamically capturing modality correlations through cross-modal attention to enhance knowledge fusion flexibility; and constructing a generalizable model adaptable to diverse intelligent education visual task scenarios, thereby improving applicability and scalability. The findings provide an effective method for deep collaboration between visual tasks and knowledge graphs in intelligent education, with significant theoretical and practical value.

1. INTRODUCTION

With the deep integration of information technology and education, intelligent education [1-3] has become the core driving force for promoting educational transformation. In intelligent education scenarios, there exists a large amount of multimodal data, including text knowledge points in teaching materials, image resources in teaching courseware, video frames of experimental processes, etc [4, 5]. These data carry rich educational knowledge. As a structured knowledge representation form, the knowledge graph can effectively organize and associate multimodal educational knowledge, providing knowledge support for visual tasks such as intelligent tutoring and personalized recommendation. However, in the process of perceiving images and other visual information, visual tasks often fail to fully utilize the textual knowledge in the knowledge graph; in the reasoning process, the knowledge graph also finds it difficult to effectively integrate the intuitive features of visual information [6-9]. The lack of synergy between the two leads to the insufficient exploitation of the value of multimodal knowledge, restricting the accuracy and intelligence level of visual tasks in intelligent education.

Carrying out research on a collaborative learning framework and optimization methods for visual tasks and knowledge graphs in intelligent education is of great significance. On the one hand, this research can break the barrier between visual information and textual knowledge in the knowledge graph, realize deep fusion of multimodal knowledge, and provide more comprehensive knowledge support for intelligent education visual tasks such as teaching image classification, knowledge point visual localization, and experimental step recognition, thereby improving the accuracy and efficiency of task processing, and assisting the optimization of educational applications such as personalized learning recommendation and intelligent question answering. On the other hand, the research results can promote the deepening of multimodal knowledge graph applications in the field of intelligent education, provide new paths for the efficient utilization and sharing of educational knowledge, promote accurate matching of educational resources, and upgrade the intelligence level of educational services.

Existing studies have made many explorations in combining multimodal knowledge processing with visual tasks in intelligent education, but there are still obvious deficiencies and shortcomings. Some studies fail to effectively solve the problem of semantic inconsistency between different modalities when processing multimodal information. For example, a simple concatenation method is used to fuse image and text features, ignoring the semantic differences between the two, which leads to a decrease in the accuracy of feature representation [10, 11]. Some studies attempt to extract association information between modalities, but the methods adopted lack flexibility. For example, the fixed-weight fusion strategy cannot dynamically capture the associated and complementary information between modalities according to different visual tasks, which limits the effect of knowledge fusion [12, 13]. In addition, most of the constructed learning frameworks have strong task specificity but lack generality and scalability. For example, the framework proposed by Arshad et al. [14] and Badrouni et al. [15] is only applicable to image classification tasks in specific disciplines and is difficult to adapt to diverse visual task scenarios in intelligent education, which is not conducive to practical application and promotion.

In response to the above problems, this paper carries out three main research tasks: first, based on contrastive learning technology, to reduce semantic inconsistency between different modalities, enhance the discriminative ability of different modality embeddings of the same entity, and achieve feature enhancement; second, to construct a cross-modal attention module to extract associated and complementary information between modalities, and to optimize text features with image features; third, to construct a general and scalable collaborative learning framework for intelligent education visual tasks and knowledge graphs, which integrates the prediction results of each modality through joint decision-making, thereby improving the accuracy of link prediction in intelligent education visual tasks. These research works not only solve the shortcomings of existing studies in terms of multimodal semantic consistency, flexibility of association information extraction, and framework generality, but also provide effective methods for deep collaboration between visual tasks and knowledge graphs in the field of intelligent education, having important theoretical value and practical significance.

2. METHOD FRAMEWORK

The teaching resources involved in intelligent education include not only text modality information such as textbook texts and courseware texts, but also image modality information such as teaching pictures and experimental video frames. These pieces of information are associated around knowledge points but have semantic differences. For example, the description of the steps of a physics experiment in text and the spatial layout of equipment in the experimental images present knowledge from logical and visual perspectives respectively, and a single modality is difficult to fully depict the connotation of knowledge. At the same time, as a structured knowledge carrier [16, 17], the knowledge graph needs to integrate multimodal information to effectively guide visual tasks, and the semantic inconsistency between different modalities may cause knowledge conflicts [18]. Therefore, a dedicated framework is needed to coordinate modality relationships, fully mine complementary information, and support the bidirectional collaboration of "visual perception–knowledge reasoning". Figure 1 shows examples of intelligent education visual tasks.

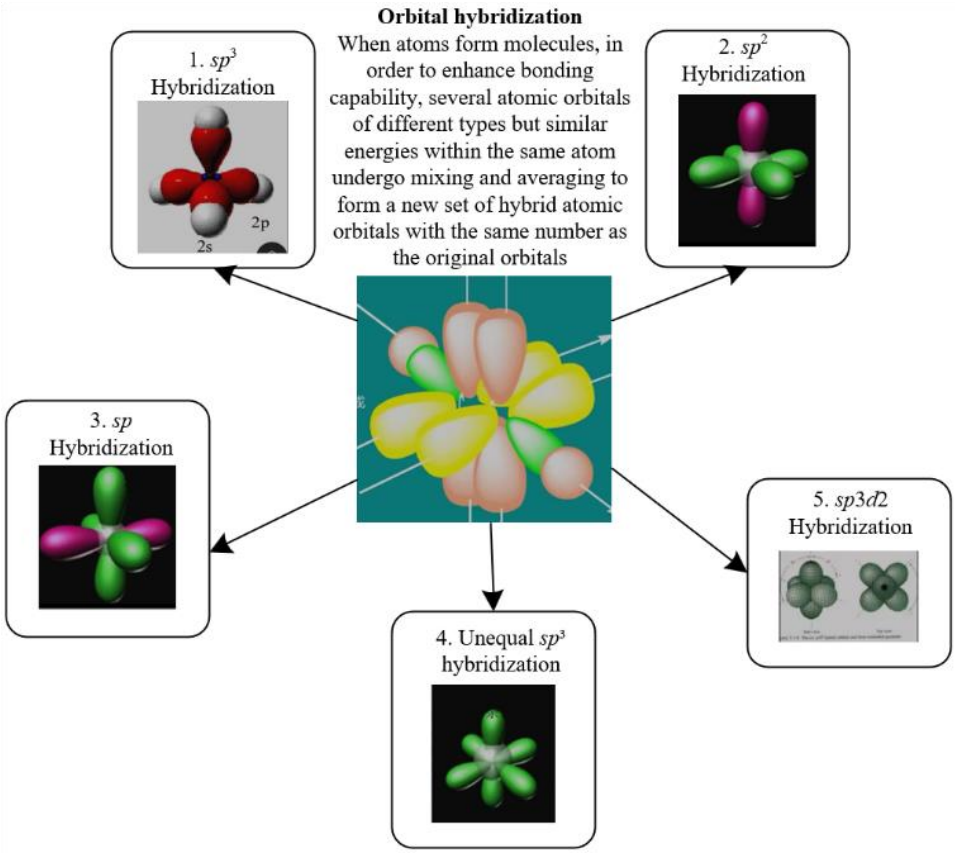


Figure 1. Examples of intelligent education visual tasks

To this end, this paper proposes a collaborative learning framework for visual tasks and knowledge graphs in intelligent education that integrates image and text information, carrying out three specific aspects of work. First, reducing modality semantic inconsistency based on contrastive learning. In intelligent education, the textual definition and image example of the same entity may have semantic deviations. By using contrastive learning to enhance the discriminative ability of different modality embeddings of the same entity, the consistency of feature representation can be improved, laying a foundation for subsequent knowledge integration and directly serving the goal of "comprehensively extracting useful knowledge". Second, constructing a cross-modal attention module. In practical applications, for example, when identifying biological specimens in teaching images, image features are more critical, while when understanding the knowledge points corresponding to the specimens, text features are more important. The attention mechanism can adaptively capture modality associations, and optimizing text features with image features can enhance knowledge complementarity, helping to accurately extract associated information. Third, constructing a general and scalable framework. By integrating multimodal prediction results through joint decision-making, the accuracy of bidirectional collaboration is improved, and at the same time, the general architecture facilitates the inclusion of new modalities or the expansion of tasks in the future, which is in line with the research goal of "constructing a closed-loop collaborative mechanism".

The proposed framework, centering on the research goal of "visual perception–knowledge reasoning" bidirectional collaboration, achieves efficient collaboration through the organic connection of three parts: multimodal information embedding, fusion, and decision-making. In the multimodal information embedding stage, the framework uses pre-trained models to extract visual features from image data and semantic features from text data respectively, and introduces contrastive learning technology. By enhancing the discriminability of the embeddings of the same educational entity in two modalities, it effectively alleviates the problem of semantic inconsistency between modalities, laying a high-quality feature foundation for subsequent fusion, and directly serving the goal of "comprehensively extracting useful knowledge." The fusion stage focuses on mining the deep association between entity text and image modalities, constructing a cross-modal attention module, and dynamically adjusting text features according to visual features, so that text information more accurately responds to visual cues in images. This strengthens the complementary association between modalities and helps the knowledge graph reasoning process better fit the results of visual perception.

The decision-making stage focuses on fully releasing the collaborative value of multimodal information, adopting a joint decision-making method to quantify the contribution weights of the image modality and text modality in the bidirectional tasks of "visual perception–knowledge reasoning". For example, in the task of knowledge point localization in teaching images, the identification results of visual regions based on image features and the reasoning conclusions of the knowledge graph corresponding to text features are both referenced. The final decision is formed through weighted integration, thereby improving the accuracy and reliability of the collaboration between visual tasks and knowledge graphs in intelligent education scenarios, and realizing full-chain bidirectional collaboration from the feature layer to the decision-making layer. Figure 2 shows the complete collaborative learning framework for intelligent education visual tasks and knowledge graphs.

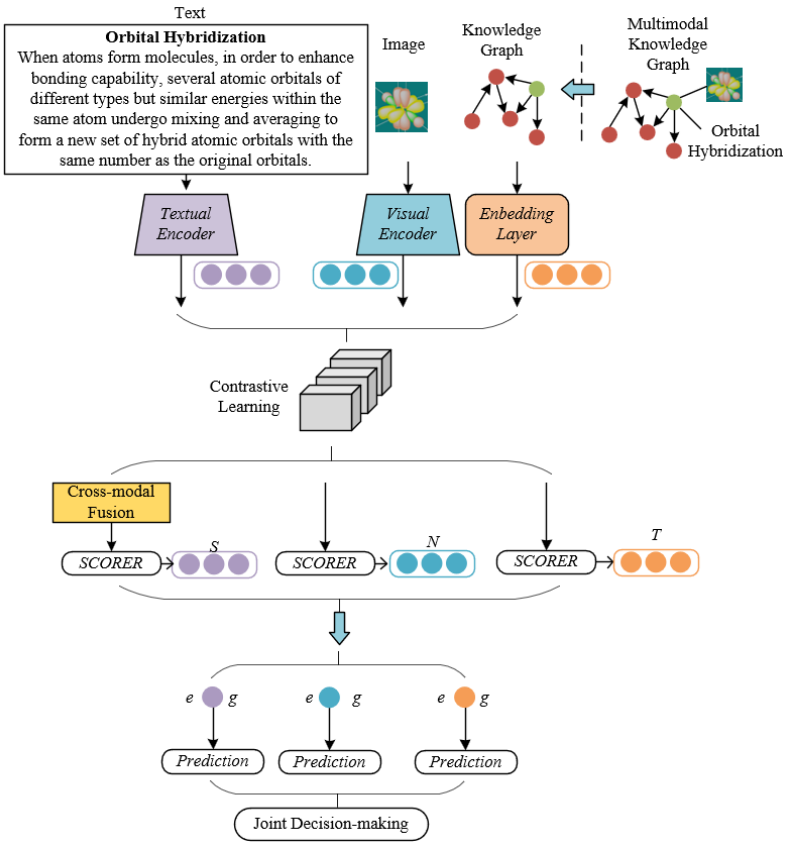


Figure 2. Collaborative learning framework for intelligent education visual tasks and knowledge graphs

2.1 Problem definition

The collaborative learning problem of visual tasks and knowledge graphs integrating image and text information studied in this paper can be defined as follows: In the multimodal knowledge graph $LJH=(R,E,F,N,S)$ in intelligent education scenarios, R is the set of educational entities, such as knowledge points, teaching resources, concepts, etc. E is the set of educational relations between entities, such as inclusion, prerequisite, association, etc. F is the set of relational triples $\{(g,e,s)\}$. N is the set of teaching images corresponding to entities, such as formula diagrams, experimental schematic diagrams, etc. S is the set of text descriptions corresponding to entities, such as concept definitions, principle explanations, etc. The goal of this problem is to perceive the images in N through visual tasks, such as entity visual localization and image content understanding, and to reason about the triples in F through the knowledge graph, so as to realize the bidirectional collaboration of "visual perception–knowledge reasoning". On the one hand, image features are used to optimize the representation of text features, enhancing the consistency between the semantic description of an entity in S and its visual presentation in N, and alleviating semantic inconsistency between modalities. On the other hand, through cross-modal association mining, visual perception results provide visual cues for knowledge graph reasoning, such as the identification of apparatus entities in experimental images assisting the reasoning of the "use" relation, while the relational knowledge in the knowledge graph guides visual tasks to focus more precisely on key entities. Finally, in the knowledge reasoning task, given (g,e), the score of the triple (g,e,s) is calculated by integrating the features and association information of the image modality and the text modality, improving the accuracy of visual task execution in intelligent education scenarios, and realizing the bidirectional empowerment of knowledge graph reasoning and visual tasks by multimodal information.

2.2 Multimodal information embedding and enhancement

Before multimodal information embedding and enhancement, this paper first extracts and initializes multimodal features. In the image feature extraction stage, the ResNet_50 model is used with targeted adjustments. Teaching images in intelligent education scenarios contain rich visual knowledge, which has essential differences from the structured information in the knowledge graph, and needs to be effectively transformed into representations that can be fused with text and structural features. ResNet_50, with its deep residual network structure, can capture multi-level features from local details to global semantics in images, meeting the needs of extracting features of complex entities in teaching images. The unification of image size is to eliminate the impact of input scale differences on feature consistency, while removing the last Softmax layer is to avoid feature compression oriented to classification tasks, retaining more original and continuous 2048-dimensional feature vectors. For text feature extraction, the BERT-base model is used to obtain a 768-dimensional pooled output. Its principle is rooted in the complexity and semantic depth of text information in multimodal knowledge graphs for intelligent education. The text corresponding to an entity not only contains literal information, but also involves contextual associations and the connotation of professional terms, which are key bases for "knowledge reasoning guiding visual perception". BERT, through its pre-trained bidirectional Transformer structure, can effectively capture the contextual semantic dependencies of text, solving problems such as polysemy and semantic ambiguity that traditional word vector models find difficult to handle, and is particularly suitable for parsing complex texts with a professional background in intelligent education. The 768-dimensional pooled output is a condensed representation of the global semantics of the text, which not only retains sufficient semantic details to support cross-modal association, but also avoids feature redundancy through dimensional control, facilitating subsequent fusion with image features. Structural feature embedding adopts the random initialization method of *TransE*. Its core principle is to stably retain the structured information of the knowledge graph in multimodal fusion, supporting efficient model convergence and the stability of "visual–knowledge" collaboration. The entity and relation structure of the knowledge graph is the backbone of "knowledge reasoning", and its embedded features need to maintain structural consistency when interacting with image and text features. The uniform distribution initialization proposed by *TransE* $(-6/(dim)^{1/2},6/(dim)^{1/2})$ controls the scale range of the initial embedding to avoid unstable model training caused by excessively large or small initial values, and is especially suitable for the complex parameter optimization process in multimodal feature fusion scenarios. Here, *dim* is the feature dimension. This design ensures that structural features have a numerical scale matching other modality features at the initialization stage, providing a numerical basis for the effective interaction of structural information with visual and text information in subsequent training. The initialized embedding layer continuously retains the relational constraints between entities during training, enabling the structured knowledge of the knowledge graph to be stably integrated into the multimodal collaboration process, ensuring that the "knowledge reasoning" stage can guide visual tasks based on reliable structural information, while also providing a structured integration framework for feeding back visual perception results into knowledge graph reasoning.

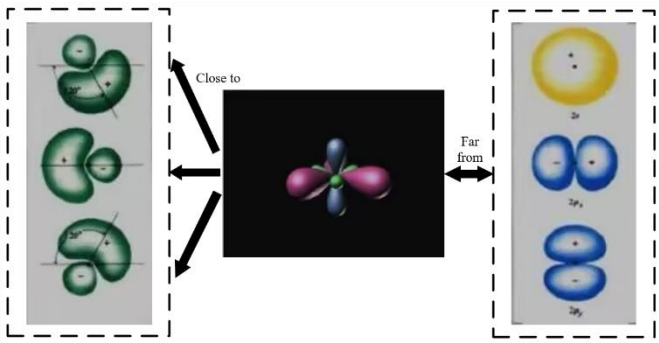


Figure 3. Example of contrastive learning

To solve the problem of modality semantic inconsistency in the "visual perception–knowledge reasoning" bidirectional collaboration of multimodal knowledge graphs in intelligent education, this paper introduces contrastive learning in the multimodal information embedding and enhancement stage to compare the similarity and difference of different modality data in a targeted manner. In intelligent education scenarios, although the image and text of the same entity point to the same knowledge, the differences in modality characteristics may lead to semantic expression deviation. Contrastive

learning, by learning the shared semantic features between modalities, can effectively weaken such deviations, making visual features and text features form a closer semantic association, and providing a basis for cross-modal knowledge transfer in bidirectional collaboration. Figure 3 shows a specific example of contrastive learning.

Entities are the basic units of knowledge organization and reasoning, and the key to "visual perception–knowledge reasoning" bidirectional collaboration lies in the precise identification of entities and the relationships between entities. If negative samples focus on modality differences, it may cause the model to mistakenly judge the features of the same entity in different modalities as unrelated, which violates the collaboration goal; whereas taking the same modality representations of different entities as negative samples can strengthen the model's perception of differences between entities, for example, distinguishing between the textual description or image features of "triangle" and "quadrilateral", ensuring that the model can accurately identify entity identity during collaboration and avoid the impact of entity confusion on the accuracy of reasoning and perception. That is, the selection of negative samples focuses on the differences between entities rather than between modalities, and its principle originates from the core demand of the multimodal knowledge graph in intelligent education.

In intelligent education, although the experimental image and textual explanation of the same knowledge point differ in form, their core semantics are consistent. The setting of positive sample pairs can guide the model to learn this intrinsic association, so that image features can automatically associate with the definition of the knowledge point in the text, and textual features can also echo the visual elements in the image, thus enabling rapid association with entity information in the knowledge graph during "visual perception", and combining visual features to refine reasoning bases during "knowledge reasoning". This paper chooses to take the representation pairs of the same entity in different modalities as positive samples, that is, by reducing the feature distance of the same entity in different modalities to strengthen semantic consistency at the entity level. Suppose the Hadamard product operation is represented by *, o, $w \in L$, $L=\{(T, N), (T, S), (N, S)\}$, and the embedding of modality o is represented by r_o. The similarity calculation formula of the positive sample pair is as follows:

$$POS_{o,w} = \sum r_o * r_w \tag{1}$$

When calculating the similarity of negative sample pairs, the diagonal elements are subtracted, the principle of which is to eliminate the interference of self-similarity of the same entity embedding in different modalities, ensuring that the negative sample pairs truly reflect the differences between different entities. In the multimodal feature matrix, diagonal elements may correspond to the embeddings of the same entity in different modalities. If they are not removed, it would cause the negative samples to contain associations that should actually be positive samples, interfering with the model's learning of entity differences. This processing ensures the purity of the negative sample set, enabling the model to learn the discriminability between different entities more accurately and improving the entity identification of feature embeddings. Let $l=\{T,N,TS\}$, the formula is as follows:

$$NEG_l = r_l \cdot r_l^S - DIAG\left(r_l \cdot r_l^S\right) \tag{2}$$

The loss calculation aims at "constraining the distance of negative samples to be greater than that of positive samples", and its principle is to guide the model to optimize the feature space distribution through quantitative constraints, ultimately serving the accuracy of the "visual perception–knowledge reasoning" bidirectional collaboration. When the positive sample distance is smaller than the negative sample distance, in the feature space learned by the model, the multimodal representations of the same entity are more clustered, and the representations of different entities are more dispersed. This distribution allows the visual task to quickly match the corresponding entity textual information in the knowledge graph when perceiving images, and also allows knowledge reasoning to effectively combine visual feature details when using textual features, thereby enhancing the reliability and accuracy of bidirectional collaboration. Suppose the hyperparameter controlling the margin distance between positive and negative sample pairs in contrastive learning is denoted by ε, and the Rectified Linear Unit is represented by $RELU(a)$ function, that is, when $a > 0$ it is a, otherwise it is 0. The loss calculation is as follows:

$$POSITIVE_{LO} = \sum_{o,w \in L} RELU\left(\varepsilon + POS_{o,w}\right) \tag{3}$$

$$NEGTIVE_{LO} = \frac{\left(\sum_{l=\{T,N,S\}} MEAN\left(RELU\left(\varepsilon - NEG_l\right)\right)\right)}{3} \tag{4}$$

$$LOSS_{CO} = POSITIVE_{LO} + NEGTIVE_{LO} \tag{5}$$

2.3 Fusion part

In intelligent education scenarios, there are complex semantic associations between the visual features of teaching images and their corresponding textual descriptions. The self-attention mechanism can enable textual features to dynamically focus on visual regions in the image related to knowledge points, while allowing image features to echo key concepts in the text. The feedforward neural network can then enhance the expression ability of this association through nonlinear transformation, adapting to the multidimensional and multilevel semantic dependencies in educational knowledge. Through the synergy of self-attention mechanism and feedforward neural network, this paper breaks the local dependency limitation of traditional sequence models, enabling global modeling of associations between multimodal features, and providing a flexible feature interaction path for transforming visual perception results into knowledge reasoning and for knowledge reasoning to guide visual tasks.

The Transformer encoder adopts multi-module stacking, and each module contains a sublayer structure with residual connection and layer normalization. The principle lies in ensuring that multimodal features retain the integrity of original information in the deep fusion process through modular design and stability mechanisms, while deepening semantic associations layer by layer, adapting to the complexity and hierarchy of multimodal data in intelligent education. The multimodal features of intelligent education often contain multi-layer information from low level to high level, and stacking multiple identical modules can achieve progressive abstraction of features, elevating the fused

features gradually from the sensory level to the semantic level. Residual connections can prevent feature degradation in deep networks, ensuring that original visual features and textual features are not diluted during the fusion process; layer normalization can standardize the input distribution of each layer, solving the training instability problem caused by the scale differences of multimodal features, ensuring robustness when the model processes diverse teaching resources, and providing a stable feature transmission channel for bidirectional collaboration. Suppose the function to be realized by each sublayer is represented by $SUBLAYER(\cdot)$, and layer normalization is represented by $LAYERNOM(\cdot)$, the output formula of a sublayer is as follows:

$$LAYERNORM\left(a+SUBLAYER\left(a\right)\right) \tag{6}$$

The attention mechanism realizes feature weighted fusion through the interaction calculation of query vector Q, key vector K, and value vector V, the principle of which is to dynamically assign attention weights to highlight key associations between modalities, so that the fused features can precisely serve the collaborative needs of intelligent education "visual perception–knowledge reasoning". In the fusion stage, textual features can be taken as Q, and image features as K and V. By calculating the similarity between Q and K, attention scores are obtained, then converted into weights through Softmax, and finally used to weight-fuse the features of V. This mechanism allows the model to adaptively focus on the most critical cross-modal associations for the current task. For example, in the "mathematical formula derivation" task, the formula symbols in the text preferentially associate with the corresponding symbol graphics in the image, thereby enhancing the task relevance of the fused features and improving the accuracy of bidirectional collaboration. Suppose the same input is represented by K, Q, and V, the activation function is represented by $SOFTMAX(\cdot)$, and the dimension number of K is represented by f_j, the formula is as follows:

$$ATTENTION\left(Q,K,V\right)=SOFTMAX\left(\frac{QK^{S}}{\sqrt{f_j}}\right)V \tag{7}$$

In intelligent education, the textual description and image presentation of the same entity have strong semantic binding, but the traditional attention mechanism may be limited to association mining within a single modality. The cross-modal attention module constructed in this paper solves the problem of poor multimodal feature interaction by switching between the text modality Q and the image modality V of the same entity, building an efficient information transmission bridge for "visual perception–knowledge reasoning" bidirectional collaboration. Figure 4 shows the architecture of the cross-modal attention module. Taking textual features as Q and image features as V enables the semantic orientation of the text to actively query the corresponding visual regions in the image, while the visual details of the image are fed back to the textual features through V, so that the text semantics obtain concrete support. This modality switching mechanism ensures the targeting of cross-modal information transmission. For example, in the "physics formula derivation" task, the symbol logic of the text formula can be precisely associated with the symbol arrangement of the formula in the image, so that visual perception results can quickly map to the logical chain of knowledge reasoning, and conversely, knowledge reasoning can also constrain the interpretation direction of visual features through textual semantics.

The multimodal data of intelligent education have significant heterogeneity, and in deep networks, it is easy to encounter gradient vanishing or explosion, leading to fusion failure. To further ensure the effective transmission of multimodal features in deep fusion through stability mechanisms, the cross-modal attention module introduces residual connection and normalization processing after calculating attention scores. Residual connection directly transmits the original features, avoiding feature degradation in deep networks, ensuring that the core knowledge point definitions in the text and the key visual elements in the image are not diluted during the fusion process; layer normalization standardizes the feature distribution of each layer's input, eliminating the interference caused by differences in scale and distribution between image and text features, enabling the model to maintain stable convergence when processing multimodal data from different disciplines, ensuring the consistency and reliability of feature transmission in bidirectional collaboration.

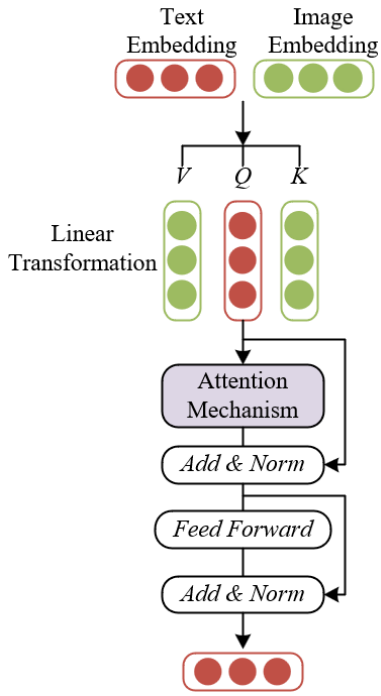


Figure 4. Architecture of the cross-modal attention module

In intelligent education scenarios, the local details of multimodal features are crucial to task completion. For example, in the text "the site of photosynthesis is the chloroplast", the qualifier "chloroplast" and the morphological details of chloroplasts inside leaf cells in the image both belong to key local features affecting knowledge reasoning and visual localization. To enhance the model's ability to model local multimodal features, the cross-modal attention module introduces a feedforward network (FFN) for nonlinear transformation. The FFN, through nonlinear transformation, can deeply process the fused features weighted by attention, strengthening the expression ability of these local features, highlighting the semantic weight of professional terms in the text, and amplifying the discriminability of key visual details in the image.

2.4 Decision part

In the decision stage, the embeddings of each modality are scored separately, and the core principle is to retain the unique knowledge of each specific modality to ensure that the irreplaceable value of images and texts in the "visual perception–knowledge reasoning" bidirectional collaboration is not submerged. In the intelligent education scenario, the image modality carries intuitive visual knowledge, and the text modality contains abstract semantic knowledge. The two play complementary roles in bidirectional collaboration. Visual perception relies on the detailed features of images, and knowledge reasoning relies on the semantic associations of text. If multimodal features are directly fused and then scored uniformly, it may lead to the dilution of key information from one modality. Scoring separately allows the knowledge of each modality to participate in decision-making independently, ensuring that image features support visual tasks and text features guide knowledge reasoning, thus providing more comprehensive modality information support for bidirectional collaboration.

The translation model *TransE* models the association between entities through the geometric relationship "$g+e\approx s$", which is highly consistent with the structural characteristics of entity relationships in the knowledge graph, and the energy function is the numerical expression of this relationship — the lower the energy, the higher the rationality of the triple (g, e, s). In intelligent education, this quantitative method can accurately describe the association strength of "visual entity–relation–text entity", facilitating the comparison between visual perception results and knowledge reasoning conclusions, and providing a unified decision scale for bidirectional collaboration. In addition, the scoring method supports replacement with other models, allowing flexible adaptation to different educational scenarios and enhancing the generality of the framework. The energy scoring function for a single modality l based on the *TransE* model is calculated as follows:

$$SCORE_l\left(g,e,s\right)=\left\|g+e-s\right\|_2^2 \tag{8}$$

The *TransE* model calculates the distance between $h+r$ and t as the energy value, which is suitable for handling simple one-to-one relationships in intelligent education; while the *TransH* model calculates by projecting entities onto a relation-specific hyperplane, which is more suitable for handling one-to-many, many-to-one, and other complex relationships. This differentiated design enables the scoring function to precisely match the diverse entity association patterns in intelligent education, ensuring reasonable scoring for the image modality when perceiving complex scenes and for the text modality when reasoning multi-level relationships, thus providing accurate numerical evidence for bidirectional collaboration. Let the normal vector of the hyperplane be denoted by q_e, and the energy function calculation formula for a single modality l based on the *TransH* model is as follows:

$$SCORE_l\left(g,e,s\right)=\left\|g-q_e^S gq_e+e-\left(s-q_e^S sq_e\right)\right\|_2^2 \tag{9}$$

The model is trained by the energy function scores of positive and negative triples, and the loss function idea is similar to *TransE*. The principle is to enhance correct associations and suppress incorrect associations, thereby improving the decision accuracy of "visual perception–knowledge reasoning" bidirectional collaboration. A positive triple corresponds to an actually existing association, and a negative triple is a destroyed incorrect association. The loss function minimizes the energy value of positive triples and maximizes the energy value of negative triples, guiding the model to learn to distinguish valid associations from invalid ones. Let the positive sample training set be denoted by F, and the negative sample training set be denoted by F', i.e., $F'=\{(g',e,s)|g'\in R\}iA(g,e,s')|s'\in R\}$, $MARGIN$ is the margin value, then the loss function is:

$$LOSS_l=\sum_{(g,e,s)\in F}\sum_{(g',e,s')\in F'}\begin{bmatrix}MARGIN\\+SCORE_l\left(g,e,s\right)\\-SCORE_l\left(g',r,s'\right)\end{bmatrix} \tag{10}$$

Let the loss value of modality l be denoted by $LOSS_l$, the loss of contrastive learning be denoted by $LOSS_{CONTRS}$, and a hyperparameter be denoted by η. The overall loss function is constructed as:

$$LOSS=LOSS_T+LOSS_N+LOSS_S+\eta LOSS_{CONTRA} \tag{11}$$

In decision-making, a joint decision method combining the decision results of each modality is adopted:

$$LOSS_b=\frac{1}{3}\left(LOSS_T+LOSS_N+LOSS_S\right) \tag{12}$$

3. METHOD OPTIMIZATION

For different educational tasks, a dynamic weight mechanism is introduced. Through reinforcement learning or meta-learning methods, the model can adjust in real time the contribution weights of image and text features according to the task type, data distribution, and entity characteristics. In the contrastive learning stage, a task-aware negative sample selection strategy is designed — for knowledge-point-dense tasks, negative samples with large text semantic differences are added; for visual recognition tasks, negative samples with large image detail differences are added. In the cross-modal attention module, scenario-aware attention bias is embedded so that in experimental teaching scenarios, priority is given to the operational step regions in the image, and in theoretical teaching scenarios, priority is given to the logical derivation parts in the text. At the same time, the feature enhancement method is optimized. Based on the domain characteristics of intelligent education data, domain pretraining weights are introduced into *ResNet* and *BERT*, reducing the adaptation cost of general models in educational scenarios and improving the domain relevance of features.

This paper constructs a deepened bidirectional interaction mechanism between the knowledge graph and visual tasks to strengthen the depth and robustness of "perception–reasoning" collaboration. On one hand, for the limitation of knowledge graph reasoning in utilizing visual features, a knowledge distillation strategy can be introduced, extracting structured knowledge such as the hierarchical relationships and causal logic of entities in the knowledge graph into constraint signals for visual features. For example, in the image feature

extraction stage, a knowledge-guided attention mask is added to make experimental image features more focused on key regions related to the knowledge point. On the other hand, the scoring mechanism in the decision stage is optimized. Combining the gradual nature of knowledge in intelligent education, knowledge temporal weights are integrated into the *TransE*/*TransH* energy function so that the association scoring between newly learned knowledge points and existing knowledge conforms better to cognitive laws. In addition, the scalability of the framework is enhanced. Through modular design, rapid access to new modalities is supported, and a transfer learning mechanism is introduced so that the framework can transfer knowledge from already learned subject tasks to new subjects, reducing the sample requirements for cross-domain adaptation and improving the landing capability in diverse intelligent education scenarios.

4. EXPERIMENTAL RESULTS AND ANALYSIS

The experiments in this paper are carried out around the "Collaborative Learning Framework and Optimization Method for Visual Tasks and Knowledge Graphs Oriented to Intelligent Education". In terms of dataset statistics, the training set and validation set have the characteristic of "validation set having larger scale and denser associations" in terms of total number of entities, total number of relations, and triple distribution. The average number of triples in the validation set is much higher than that in the training set (see Table 1), that is, the validation set covers more complex multimodal entity associations, which provides sufficient multimodal samples for feature enhancement in contrastive learning, and also creates a data foundation for the cross-modal attention module to mine association complementary information. In terms of parameter settings, 1500 epochs ensure deep learning of multimodal features and the knowledge graph structure by the model. A learning rate of 0.001 and weight decay of 0.001 work together to suppress overfitting, adapting to the stable convergence of feature enhancement in contrastive learning; a dropout of 0.2 further enhances the robustness of the model, while the early stop strategy with *patience*=10 and *EarlyStop*=5 avoids insufficient training while preventing overfitting (see Table 2).

Table 1. Dataset statistical information table

	Training Set	Validation Set
Total number of entities	11256	13256
Total number of relations	265	224
Training triples	68595	265844
Validation triples	18956	16589
Test triples	9852	21425
Total number of triples	98625	325625
Average number of triples	7.24	22.32

Table 2. Parameter settings

Parameter	Fixed Value
Epochs	1500
Learning Rate	0.001
Dropout	0.2
Weight Decay	0.001
Patience	10
Early Stop	5

Table 3. Comparative experiment

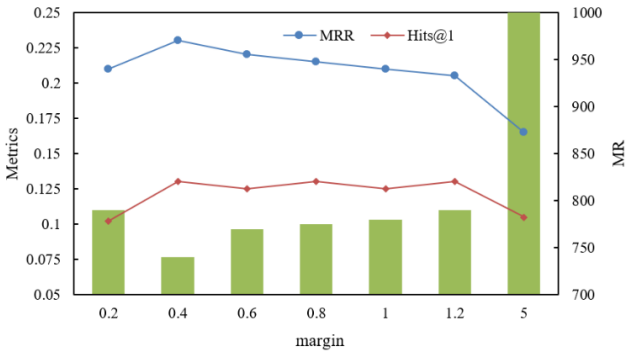
Model	Training Set					Validation Set				
	MR	*MRR*	*Hits*@1	*Hits*@10	*Hits*@100	*MR*	*MRR*	*Hits*@1	*Hits*@10	*Hits*@100
TransE	1253.35	0.162	8.62	31.25	51.23	425.23	0.156	11.25	27.54	57.23
TransR	1235.25	0.112	7.56	21.23	42.23	475.23	0.132	9.56	27.23	56.23
TransH	1325.23	0.165	10.25	25.36	44.56	654.23	0.125	6.32	32.15	47.52
TransD	1256.23	0.114	7.89	21.32	42.23	478.23	0.132	9.45	22.36	53.23
Transparse	987.23	0.148	7.62	31.25	52.36	416.25	0.164	12.36	28.23	57.23
MTransE	915.23	0.135	6.54	32.32	54.23	378.23	0.168	11.25	30.21	62.35
CTransR	846.23	0.159	9.56	31.25	56.36	345.23	0.165	12.35	31.25	62.34
Proposed model (*TransE*)	735.23	0.215	11.23	42.56	63.25	325.32	0.178	12.23	33.32	63.34
Proposed model (*TransH*)	865.23	0.166	11.23	33.54	58.23	412.23	0.168	11.28	33.54	63.25

Table 4. Ablation experiments

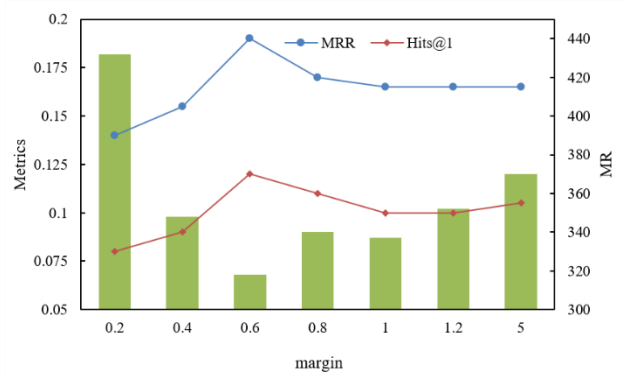
Model	Training Set					Validation Set				
	MR	*MRR*	*Hits*@1	*Hits*@10	*Hits*@100	*MR*	*MRR*	*Hits*@1	*Hits*@10	*Hits*@100
Remove Text Modality	965.23	0.156	9.45	31.26	51.23	375.23	0.162	11.23	28.56	58.23
Remove Contrastive Learning	835.23	0.182	11.23	35.62	58.62	345.23	0.168	12.23	32.23	62.31
Remove Fusion Module	754.23	0.215	12.23	37.52	61.23	326.31	0.175	11.25	31.25	62.32
Remove Image Modality	756.23	0.216	12.25	37.56	61.25	336.23	0.174	12.23	32.25	62.54
Complete Model	735.23	0.219	13.24	42.36	63.23	325.26	0.178	12.29	32.36	64.58

From the comparative experimental results shown in Table 3, the model in this paper shows significant advantages over traditional Trans series models such as *TransE*, *TransR*, *TransH*, *TransD* in the *MR*, *MRR*, and *Hits*@1/10/100 indicators, mainly due to the collaborative effect of the three innovative modules of the research work. In traditional models, the *MR* of the *TransE* training set is 1253.35, while the model in this paper based on *TransE* is reduced to 735.23, and in the validation set it is reduced from 425.23 to 325.32. The significant decrease in *MR* indicates that contrastive learning effectively reduces multimodal semantic inconsistency, making entity and relation embeddings more accurate and reducing the average ranking of prediction results. At the same time, the improvement of the *MRR* indicator reflects that contrastive learning enhances the discrimination ability of "multimodal embeddings of the same entity", allowing the model to more clearly identify semantic associations, such as the *MRR* of the TransE training set in this paper reaching 0.215,

compared to only 0.162 in the traditional model; in the validation set it increases from 0.156 to 0.178. The progress in the *Hits* series indicators is particularly prominent. Taking *Hits*@10 as an example, in the *TransE* training set of this paper it reaches 42.56, while the traditional *TransE* is only 31.25; in the validation set it increases from 27.54 to 33.32. Based on *TransH*, the *Hits*@10 in the training set of this paper increases from 25.36 to 33.54, and in the validation set from 32.15 to 33.54. This shows that the cross-modal attention module successfully extracts the association complementary information between images and texts, optimizes the semantic consistency of feature representation, makes the model easier to capture correct association relationships, and improves the hit rate of *Top-K* prediction. The consistent improvement of the model in this paper based on *TransE* and *TransH*, such as the *MR* of the *TransH* version in the training set being reduced from 1325.23 to 865.23, and in the validation set from 654.23 to 412.23, verifies the universality and scalability of the framework. After the joint decision mechanism integrates multimodal prediction results, not only is a breakthrough in accuracy achieved in the training set, but also a stable advantage is maintained in the validation set, indicating that multimodal interaction enhances the robustness of reasoning and adapts to the scenario requirements of "complex knowledge associations and heterogeneous modal data" in intelligent education.



(a) Training set



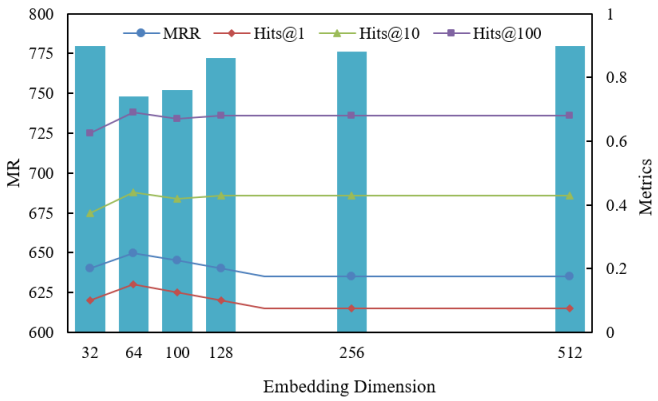
(b) Validation set

Figure 5. Effect of different margin values on performance

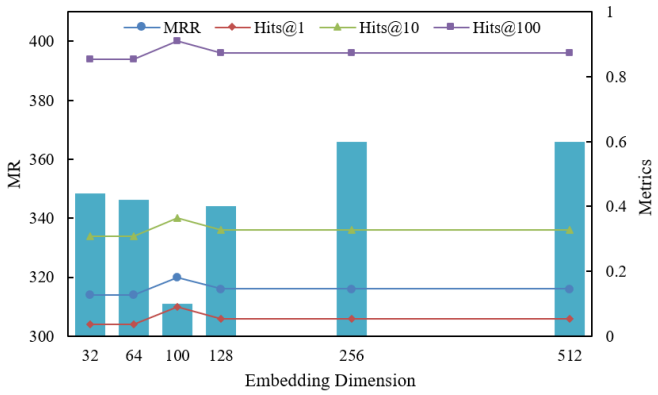
The ablation experiments, by gradually removing the core modules, clearly reveal the collaborative action mechanism of the three technical modules in the research. As can be seen from Table 4, when removing the text modality, the *MR* in the training set rises sharply from 735.23 to 965.23, and in the validation set from 325.26 to 375.23, the MRR drops from 0.219 to 0.156, and in the validation set from 0.178 to 0.162, *Hits*@1 decreases from 13.24 to 9.45. This indicates that the "knowledge logic, concept definition" carried by the text modality is the core of semantic understanding. After its absence, the semantic richness of entity embeddings drops sharply, resulting in the deterioration of prediction ranking. In contrast, when removing the image modality, the *MR* in the training set rises to 756.23, and *Hits*@10 drops from 42.36 to 37.56, indicating that the "spatial structure, visual features" provided by the image modality are a key supplement to the text, and the absence of either will damage the integrity of feature representation, verifying that multimodal integration is the basic premise of the framework. After removing contrastive learning, the *MR* in the training set rises from 735.23 to 835.23, the *MRR* drops from 0.219 to 0.182, and *Hits*@10 drops from 42.36 to 35.62, reflecting that contrastive learning, through "semantic difference constraint between modalities", effectively reduces the representation conflict between images and text, and enhances the discrimination ability of "multimodal embeddings of the same entity." Without this module, semantic inconsistency between modalities is amplified, making it difficult for the model to accurately capture association relationships, proving that contrastive learning is the core engine for achieving multimodal feature enhancement. When removing the fusion module, the *MR* in the training set rises to 754.23, and *Hits*@10 drops from 42.36 to 37.52, indicating that this module, by dynamically focusing on the "complementary regions between modalities", deeply mines the association value between visual details and semantic logic. Without it, the association information between modalities is not fully integrated, the feature complementarity cannot be released, and it highlights that cross-modal attention is the key bridge to realizing multimodal collaborative reasoning. The complete model achieves the best in all indicators: training set *MR* 735.23, *MRR* 0.219, *Hits*@10 42.36; validation set *MR* 325.26, *MRR* 0.178, *Hits*@10 64.58, confirming the progressive enhancement logic of "contrastive learning aligns semantics → fusion module mines associations → multimodal integration provides raw material": contrastive learning solves "modality heterogeneity conflict", the fusion module amplifies "complementary information value", and image and text provide "full-dimensional feature support". The collaboration of the three enables the model to achieve breakthroughs of better ranking, higher confidence, and stronger *Top-K* hit rate in intelligent education knowledge link prediction, laying the technical rationality for the application of multimodal knowledge graph in educational scenarios.

Figure 5, through the *MR*, *MRR*, and *Hits*@1 indicators of the training set and validation set, reveals the regulation mechanism of the *margin* hyperparameter in contrastive learning on the multimodal feature discrimination ability and framework generalization. In the training set, when the margin is in the range 0.2–1.2, *MRR* maintains a high level of 0.21–0.23, *Hits*@1 is stable at 0.12–0.13, and *MR* fluctuates slightly, indicating that in this range contrastive learning can effectively narrow the multimodal semantic difference while avoiding overfitting. The positive and negative sample spacing is moderate, ensuring both the discrimination degree of "multimodal embeddings of the same entity" and retaining generalization space for the joint decision framework. When the margin increases to 5, the *MR* in the training set suddenly rises to over 950, *MRR* and *Hits*@1 significantly decrease, reflecting that an excessively large spacing forces the model to extremely distinguish positive and negative samples in the

training set, leading to feature learning deviating from real semantic associations and causing overfitting. In the validation set, when *margin* = 0.6, *MRR* reaches the peak, *MR* drops to the lowest, and *Hits*@1 also performs best, indicating that at this value, the multimodal features generated by contrastive learning not only accurately capture the association between "image details and text knowledge" in educational scenarios but also, through reasonable positive and negative sample constraints, enable the joint decision framework to stably integrate multimodal prediction results. In summary, the optimization of the *margin* needs to balance between "multimodal semantic discrimination" and "model generalization": too low and the modality heterogeneity is not sufficiently constrained, too high and the training bias is amplified; while values around 0.6 just support the collaboration of the three main research modules of the paper—contrastive learning achieves precise feature alignment, cross-modal attention effectively mines associations, and joint decision integrates results—ultimately improving the accuracy and robustness of intelligent education knowledge link prediction.



(a) Test set



(b) Training set

Figure 6. Effect of different embedding dimensions on performance

Figure 6, through the *MR*, *MRR*, *Hits*@1/10/100 indicators of the training set and test set, reveals the regulation law of embedding dimension on multimodal feature expression ability, cross-modal association mining, and model generalization, which is deeply related to the collaborative mechanism of the three main research modules of the paper. In the training set, when the embedding dimension increases from 32 to 100, *MR* decreases significantly, *MRR* climbs, and *Hits*@10 increases synchronously, indicating that a moderate dimension provides sufficient semantic discrimination space for contrastive learning: the visual features of images and the knowledge descriptions of text can be more accurately aligned at this dimension, while supporting the cross-modal attention module to mine the association complementary information from "visual details → semantic logic". However, when the dimension exceeds 100, *MR* in the training set starts to rise again, *MRR* and the *Hits* series indicators fall back, reflecting that at high dimensions the model overfits the detail noise of multimodal data, causing the "semantic alignment" of contrastive learning to degenerate into "noise memorization", and the cross-modal attention also falls into mining invalid associations, destroying the effectiveness of feature enhancement. In the test set, when the dimension increases from 32 to 100, *MR* decreases from over 750 to over 730, *MRR* increases from 0.21 to 0.23, and *Hits*@10 increases from 0.45 to 0.47, verifying the "Goldilocks zone" around 100 dimensions: at this point multimodal features retain enough semantic discrimination without introducing a generalization bottleneck due to excessive dimensions. But when the dimension exceeds 100, *MR* in the test set rises sharply to over 750, and *MRR* and *Hits* indicators drop sharply, revealing that at high dimensions the multimodal heterogeneity is amplified: the visual signal of images and the symbolic semantics of text are more difficult to align in high-dimensional space, and the cross-modal attention cannot effectively capture core associations, ultimately leading to a collapse of the generalization ability of link prediction. In summary, the optimization of embedding dimension is essentially the dynamic balance between "multimodal feature expression ability" and "model generalization": Low dimension (<100): the capacity of the feature space is insufficient, contrastive learning cannot effectively distinguish multimodal semantic differences, and cross-modal attention cannot mine deep complementary information, leading to underfitting; Moderate dimension (around 100): precisely supports the collaboration of the three modules in the paper—contrastive learning efficiently aligns multimodal semantics, cross-modal attention deeply extracts association complementary information, and the joint decision framework stably integrates prediction results, achieving the optimal balance between link prediction accuracy and generalization; High dimension (>100): the feature space becomes overly complex, the model fits the detail noise of multimodal data, semantic alignment of contrastive learning fails, cross-modal association mining deviates, ultimately causing overfitting and destroying the robustness of knowledge graph reasoning in intelligent education scenarios.

5. CONCLUSION

This paper, focusing on the collaborative problem of visual tasks and knowledge graphs in intelligent education, formed a systematic solution through three core research works: based on contrastive learning technology, it effectively reduced the semantic inconsistency between image and text modalities, enhanced the discrimination ability of multimodal embeddings of the same entity, and laid a feature foundation for cross-modal collaboration; the constructed cross-modal attention module successfully mined the association complementary information between modalities, optimized text features through image features, and strengthened the bidirectional information transmission of "visual perception–knowledge

reasoning"; on this basis, the formed general and scalable framework, by integrating multimodal prediction results through joint decision, significantly improved the accuracy of link prediction for visual tasks in intelligent education. The experimental results show that the framework outperforms traditional methods in key indicators such as *MR*, *MRR*, and *Hits@K*, verifying the collaborative effectiveness of contrastive learning, cross-modal attention, and joint decision mechanisms. The research value lies in breaking the limitation of single-modality knowledge utilization, realizing deep integration of multimodal educational knowledge, providing more accurate knowledge support for educational applications such as intelligent tutoring and personalized learning recommendation, and promoting the practical process of multimodal knowledge graphs in the field of intelligent education.

However, the research still has certain limitations: first, the current framework mainly focuses on image and text modalities, and its ability to integrate other educational modalities such as audio and animation is insufficient, making it difficult to cover the full-modality needs of complex educational scenarios; second, the generalization performance on small-sample educational data needs to be improved, and the adaptability of contrastive learning and cross-modal attention needs further optimization; third, the dynamic evolution characteristics of educational knowledge are not fully considered, and the ability of the framework to process temporal knowledge is weak. Future research can advance in three aspects: expanding the range of multimodal fusion, introducing modality-adaptive mechanisms to achieve unified representation of images, text, audio, and video; exploring meta-learning-based small-sample adaptation strategies to improve the robustness of the framework in scenarios with scarce educational data; and combining the temporal characteristics of educational knowledge to design dynamic knowledge update modules, making the framework more aligned with the evolution law of knowledge in the actual teaching process, and ultimately achieving deep coupling with intelligent education scenarios.

ACKNOWLEDGMENT

This paper is a phased achievement of two projects funded by the Fundamental Research Funds for Autonomous Region Universities of Inner Mongolia: "Application of Knowledge Graph Technology in Ideological and Political Courses in Universities" (Grant No.: JY20250061) and "Research on the Development and Evolution of Industrial Culture in Inner Mongolia" (Grant No.: ZLJD250101).

REFERENCES

[1] Sanusi, I.T., Agbo, F.J., Dada, O.A., Yunusa, A.A., Aruleba, K.D., Obaido, G., Oyelere, S.S. (2024). Stakeholders' insights on artificial intelligence education: Perspectives of teachers, students, and policymakers. Computers and Education Open, 7: 100212. https://doi.org/10.1016/j.caeo.2024.100212

[2] Olari, V., Romeike, R. (2024). Data-related concepts for artificial intelligence education in K-12. Computers and Education Open, 7: 100196. https://doi.org/10.1016/j.caeo.2024.100196

[3] Leitner, M., Greenwald, E., Wang, N., Montgomery, R., Merchant, C. (2023). Designing game-based learning for high school artificial intelligence education. International Journal of Artificial Intelligence in Education, 33(2): 384-398. https://doi.org/10.1007/s40593-022-00327-w

[4] Venkatachalam, C., Venkatachalam, S. (2024). Optimal intelligent information retrieval and reliable storage scheme for cloud environment and E-learning big data analytics. Knowledge and Information Systems, 66(11): 6643-6673. https://doi.org/10.1007/s10115-024-02152-0

[5] Kalaivanan, E., Brindha, S. (2022). Deep learning based big data analytics on traffic congestion in urban intelligent transportation system. International Journal of Early Childhood Special Education, 14(3): 9008-9010.

[6] Dong, H., Wang, P., Xiao, M., Ning, Z., Wang, P., Zhou, Y. (2024). Temporal inductive path neural network for temporal knowledge graph reasoning. Artificial Intelligence, 329: 104085. https://doi.org/10.1016/j.artint.2024.104085

[7] Zhu, H., Xu, D., Huang, Y., Jin, Z., Ding, W., Tong, J., Chong, G. (2024). Graph structure enhanced pre-training language model for knowledge graph completion. IEEE Transactions on Emerging Topics in Computational Intelligence, 8(4): 2697-2708. https://doi.org/10.1109/TETCI.2024.3372442

[8] Deng, C., Yu, Q., Luo, G., Zhao, Z., Li, Y. (2022). Big data-driven intelligent governance of college students' physical health: System and strategy. Frontiers in Public Health, 10: 924025. https://doi.org/10.3389/fpubh.2022.924025

[9] Ma, J., Pan, L. (2025). Image and text aspect-level sentiment analysis based on attentional mechanisms and bimodal fusion. International Journal of Decision Support System Technology (IJDSST), 17(1): 1-23. https://doi.org/10.4018/IJDSST.370388

[10] Zhang, Y., Han, S., Zhang, Z., Wang, J., Bi, H. (2023). CF-GAN: Cross-domain feature fusion generative adversarial network for text-to-image synthesis. The Visual Computer, 39(4): 1283-1293. https://doi.org/10.1007/s00371-022-02404-6

[11] Kadam, V.S., Pingale, S., Biradar, S.R., Rohokale, V.M., Bamane, K.D. (2025). Designing a novel framework of email spam detection using an improved heuristic algorithm and dual-scale feature fusion-based adaptive convolution neural network. Information Security Journal: A Global Perspective, 34(4): 286-309. https://doi.org/10.1080/19393555.2024.2432258

[12] Lu, S., Ding, Y., Liu, M., Yin, Z., Yin, L., Zheng, W. (2023). Multiscale feature extraction and fusion of image and text in VQA. International Journal of Computational Intelligence Systems, 16(1): 54. https://doi.org/10.1007/s44196-023-00233-6

[13] Tuncer, T., Barua, P.D., Tuncer, I., Dogan, S., Acharya, U.R. (2024). A lightweight deep convolutional neural network model for skin cancer image classification. Applied Soft Computing, 162: 111794. https://doi.org/10.1016/j.asoc.2024.111794

[14] Arshad, T., Zhang, J., Anyembe, S.C., Mehmood, A. (2024). Spectral spatial neighborhood attention transformer for hyperspectral image classification: Transformateur d'attention de voisinage spatial-spectral pour la classification d'images hyperspectrales. Canadian Journal of Remote Sensing, 50(1): 2347631.

[15] Badrouni, M., Katar, C., Inoubli, W. (2024). Large-scale

knowledge graph representation learning. Knowledge and Information Systems, 66(9): 5479-5499. https://doi.org/10.1007/s10115-024-02131-5

[16] Shokrzadeh, Z., Feizi-Derakhshi, M.R., Balafar, M.A., Mohasefi, J.B. (2024). Knowledge graph-based recommendation system enhanced by neural collaborative filtering and knowledge graph embedding. Ain Shams Engineering Journal, 15(1): 102263. https://doi.org/10.1016/j.asej.2023.102263

[17] Zhang, L., Chen, L., Zhou, C., Li, X., Yang, F., Yi, Z. (2023). Weighted graph-structured semantics constraint network for cross-modal retrieval. IEEE Transactions on Multimedia, 26: 1551-1564. https://doi.org/10.1109/TMM.2023.3282894

[18] Mbiaya, F.A., Vrain, C., Ros, F., Dao, T.B.H., Lucas, Y. (2024). Knowledge graph-based image classification. Data & Knowledge Engineering, 151: 102285. https://doi.org/10.1016/j.datak.2024.102285