MMCPL: MULTIMODAL EMBEDDING ENHANCEMENT FOR CONCEPT PREREQUISITE LEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Concept prerequisite relation learning aims to identify directed dependencies between concepts, which is crucial for personalized education and learning path planning in intelligent education. Existing methods rely heavily on text modalities to model concepts, neglecting the crucial implicit information embedded within image modalities. We propose a **M**ulti**M**odal Embedding Enhancement method for **C**oncept **P**rerequisite **L**earning (MMCPL). Specifically, MMCPL extracts textual and image modal information of concepts from open authoritative knowledge bases and utilizes pretrained models to obtain multimodal representations of these concepts. Furthermore, we employ a cross-modal attention mechanism to facilitate bidirectional information interaction between textual and image modalities. By integrating Gaussian noise based intra-modal contrastive learning with inter-modal contrastive learning, we synergistically optimize the cross-modal consistency of representations while further enhancing the discriminative power within each modality. Extensive experiments on three benchmark datasets demonstrate that MMCPL consistently outperforms competitive baseline methods across all evaluation metrics.

1 INTRODUCTION

With the rapid development of intelligent learning platforms and personalized education systems, automatically and accurately constructing knowledge structures has become a fundamental challenge in educational data mining and artificial intelligence in education (O'Dea & Stern, 2022). Prerequisite relations characterize the dependencies among concepts, specifying the fundamental concept that must be mastered prior to the acquisition of a target concept (Zhang et al., 2025a). For example, in mathematics, understanding *Triangle* typically requires prior mastery of *Angle* and *Line Segment*. Consequently, these relations serve as the foundation for downstream tasks such as learning path planning (Yu et al., 2024; Wang & Liu, 2016), cognitive diagnosis (Song et al., 2023; Gao et al., 2023), and educational resource recommendation (Dong et al., 2024; Wang et al., 2022).

In earlier studies, concept prerequisite extraction primarily relied on manual annotations by experts or traditional machine learning approaches (Talukdar & Cohen, 2012; Liang et al., 2015). Although these methods can ensure high accuracy, they are labor-intensive and difficult to scale to rapidly evolving knowledge systems. With the emergence of deep learning, researchers have shifted toward data-driven methods (Roy et al., 2019). Most of these approaches employ pretrained language models to encode the textual descriptions and infer prerequisite by capturing semantic dependencies. Some studies further utilize graph neural networks to explore structural dependencies within knowledge networks, improving automation and generalization (Sun et al., 2022; Zhang et al., 2022). However, existing research predominantly models concepts solely based on textual descriptions, remaining constrained by single-modality limitations. In fact, visual representations often encapsulate structural details difficult to exhaustively describe in text (Krishna et al., 2017; Lu et al., 2022); yet, few studies have leveraged such multimodal information to infer prerequisite relations.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

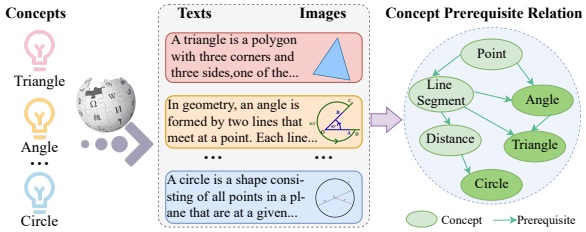

Figure 1: Leveraging multimodal information of concepts for concept prerequisite relation learning.

Although multimodal techniques have demonstrated considerable potential (Zhu et al., 2024), existing studies still suffer from two major gaps and limitations. First, the application of multimodal learning to concept prerequisite relation modeling remains scant, with the majority of existing approaches confined to uni-modal textual frameworks (Wang et al., 2024). Second, existing multimodal representation learning methods often face challenges in aligning cross-modal semantics and dealing with noise introduced by fluctuations in image quality when integrating text and images (Liang et al., 2022). As a result, the learned multimodal representations tend to exhibit weak semantic coherence and insufficient discriminative capability, which makes them inadequate for supporting accurate inference of concept prerequisite relations. Therefore, as illustrated in Figure 1, it is necessary to explore how to incorporate multimodal techniques into the task of concept prerequisite relation learning to overcome the limitations of uni-modal text modeling.

To this end, we propose a Multimodal Embedding Enhancement method for Concept Prerequisite Learning (MMCPL), designed to fully leverage concept multimodal information to construct richer concept representations, thereby enhancing the model's ability to identify prerequisite relations. Specifically, we extract textual and image modality information for concepts from the open knowledge base, Wikipedia. To achieve effective alignment between the two modalities and mitigate representation noise caused by fluctuations in image quality, we further design a cross-modal enhanced embedding module. This module employs a cross-modal attention mechanism to facilitate information interaction and strengthen shared feature extraction between modalities. Furthermore, we incorporate both intra- and inter-modal contrastive learning to enhance the distinctiveness of uni-modal representations while capturing consistency and inconsistency across modalities. This combined approach improves the discriminative power of concept representations, ultimately boosting the overall learning capability of the model.

The main contributions of this paper are as follows:

- We propose a novel multimodal embedding enhancement approach for concept prerequisite relation learning, integrating textual and image information to model multidimensional concept characteristics and predict prerequisite relations.
- We introduce a collaborative strategy that combines a cross-modal attention mechanism with dual branch contrastive learning, which effectively mitigates noise from image backgrounds and promotes information alignment across modalities.
- Extensive experiments conducted on three public datasets demonstrate that, compared with representative state-of-the-art baseline methods, MMCPL achieves superior performance.

2 Related Work

2.1 Concept Prerequisite Relation Learning

The core objective of concept prerequisite relation learning is to accurately identify directed dependencies between foundational concepts and advanced concepts, thereby providing essential support for intelligent education tasks. Talukdar and Cohen pioneered the study of extracting prerequisite relationships from Wikipedia (Talukdar & Cohen, 2012), laying the foundation for this research area. Early studies primarily relied on traditional statistical and machine learning methods. RefD (Liang et al., 2015) proposed a citation distance based measurement strategy that leveraged the hyperlink structure of Wikipedia to infer prerequisite relationships across domains. Sayyadiharikandeh et al. (Sayyadiharikandeh et al., 2019) further incorporated Wikipedia clickstream data, using learning behavior signals to assist relation identification, which expanded the range of available signals and data sources. With the development of deep learning techniques, research in this area has gradually shifted toward end-to-end modeling frameworks. PREREQ (Roy et al., 2019) adopts a Siamese

network architecture to learn semantic representations of concepts and determines prerequisite relationships between concept pairs based on these representations. With the growing prominence of graph neural networks, subsequent studies have increasingly emphasized structured graph based modeling. ConLearn (Sun et al., 2022) integrates concept graphs with self-attention mechanisms to enable context aware prerequisite relation learning. Methods such as MHAVGAE (Zhang et al., 2022), HGAPNet (Mazumder et al., 2023), and LCPRE (Sun et al., 2024) construct heterogeneous concept resource graphs that incorporate multi source entities and relations to capture more complex concept dependency patterns. GKROM (Zhang et al., 2025b) further enhances semantic representations of concepts and learning resources by introducing multiple relation types and global optimization strategies, thereby optimizing the dependency network from a holistic perspective and improving prerequisite relation identification performance.

2.2 Multimodal Representation Learning

Multimodal representation learning aims to enhance a model's ability to understand and reason about data by integrating information from multiple modalities (Li et al., 2020b). Early approaches primarily adopted fusion based or interaction based architectures, in which image region features and textual tokens are jointly fed into Transformer models to enable fine grained alignment and cross-modal reasoning. Representative methods include UNITER (Chen et al., 2020b), ViLBERT (Lu et al., 2019), and LXMERT (Tan & Bansal, 2019). Although these approaches exhibit strong cross-modal semantic modeling capabilities, they are often sensitive to the quality of paired data and incur high computational costs. Subsequently, alignment based pretraining emerged as an important research direction. CLIP (Radford et al., 2021) employs dual encoders and contrastive learning to obtain transferable cross-modal representations in a shared semantic space. BLIP (Li et al., 2022) introduces a generate and filter bootstrapping strategy to alleviate noise in web scale image text data and adopts a unified vision language pretraining paradigm that supports both understanding and generation tasks. In addition, ImgFact (Liu et al., 2024), as a multimodal knowledge graph, aligns images with factual triples for the first time, providing a new perspective for multimodal knowledge graph modeling. MMRL (Guo & Gu, 2025) proposes a framework for cross-modal interaction in a shared and modality invariant representation space, and significantly improves the generalization ability of multimodal embeddings through higher level representation learning. Despite the notable progress achieved by existing multimodal representation learning methods in semantic concept representation, they have not been specifically designed for concept prerequisite relation learning. This research gap motivates our work to incorporate multimodal concept information into the task of prerequisite relation learning.

3 Methodology

In this paper, we introduce a novel model designed to leverage multimodal embedding enhancement for concept prerequisite relation learning. Figure 2 illustrates the overall architecture of our proposed MMCPL.

3.1 Problem Statement

In this paper, we formalize concept prerequisite relation learning as a binary classification problem. Let $C = \{c_i\}_{i=1}^{m}$ denote the set of concepts and $R = \{r_j\}_{j=1}^{n}$ denote the set of learning resources, which provide rich contextual information for these concepts. m and n represent the numbers of concepts and learning resources, respectively. Given a concept pair (c_i, c_j), our objective is to predict whether concept c_i is a prerequisite of concept c_j. This task can be formulated as learning a target mapping function $F(c_i, c_j) \rightarrow \{0, 1\}$, where $F(c_i, c_j) = 1$ indicates that c_i is a prerequisite concept of c_j, and $F(c_i, c_j) = 0$ otherwise.

3.2 Concept Multimodal Representation Generation Module

For concepts without available images, we synthesize visual content using Wan2.5-t2i-preview, and the missing-image ratios vary across datasets. Generated images are filtered by semantic consistency and visual-quality criteria before encoding, and the complete retrieval, prompting, generation,

and filtering pipeline is provided in Appendix A.1. To further investigate whether synthesized concept images introduce representation noise, we conduct a comparative evaluation between real and synthetic subsets; detailed results and discussions are provided in Appendix A.2.

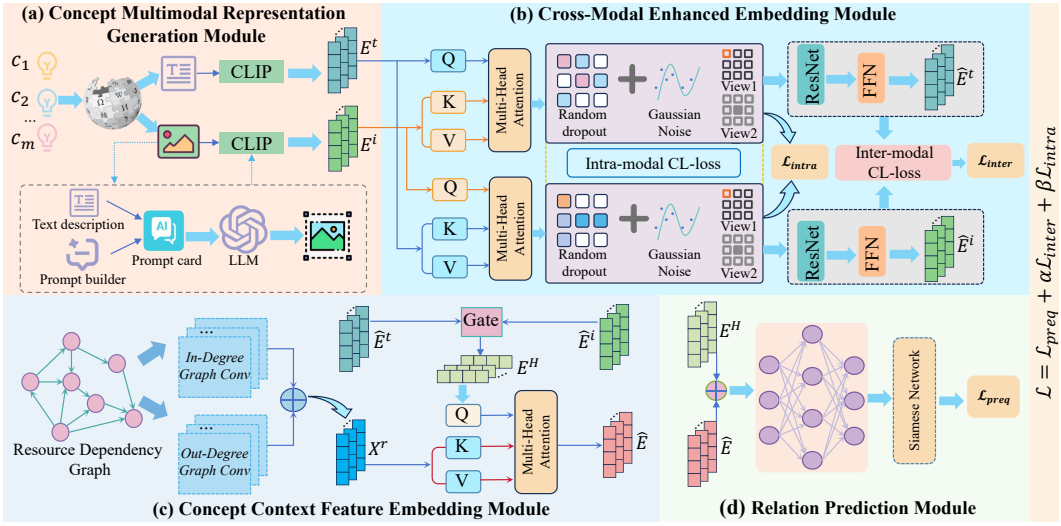

Figure 2: The overall structure of the proposed MMCPL.

3.3 Cross-Modal Enhanced Embedding Module

Although features from the CLIP encoder exhibit strong transferability, visual semantics are often entangled and contaminated by background noise (Li et al., 2025). To mitigate this, after obtaining textual and image features, we design a bidirectional cross-modal multi-head attention mechanism for feature enhancement, extracting more discriminative correlation features. Specifically, this module facilitates interaction from two perspectives. On the one hand, the initial textual embeddings $\boldsymbol{E}^t$ serve as queries, while the initial image embeddings $\boldsymbol{E}^i$ act as keys and values, enabling image information to enhance textual representations. On the other hand, $\boldsymbol{E}^i$ act as queries with $\boldsymbol{E}^t$ functioning as keys and values, allowing textual information to refine image representations. This bidirectional interaction effectively reduces the semantic gap between modalities and strengthens shared representations. The computation is formulated as follows:

$$\widetilde{\boldsymbol{E}}^t = \mathrm{MultiAtt}\left(\boldsymbol{E}^t, \boldsymbol{E}^i, \boldsymbol{E}^i\right); \quad \widetilde{\boldsymbol{E}}^i = \mathrm{MultiAtt}\left(\boldsymbol{E}^i, \boldsymbol{E}^t, \boldsymbol{E}^t\right), \tag{1}$$

where $\widetilde{\boldsymbol{E}}^t, \widetilde{\boldsymbol{E}}^i \in \mathbb{R}^{m\times d}$ denote the enhanced textual and image modality embeddings obtained after the cross-modal attention mechanism, respectively.

Although cross-modal enhanced embeddings integrate multimodal information, they may still suffer from representation overlap among semantically similar concepts. To address this, we introduce intra-modal contrastive learning for both textual and image embeddings. Specifically, for each sample, we generate two augmented views by combining random feature dropout with mild Gaussian perturbations. Views from the same concept serve as positive pairs, while those from distinct concepts serve as negatives. Optimized via the InfoNCE (Chen et al., 2020a) loss with temperature scaling, this objective improves the discriminability and robustness of the enhanced embeddings, while mitigating noise from image quality variations. The intra-modal contrastive loss for a single modality is defined as follows:

$$\mathcal{L}_{intra}(X) = -\frac{1}{2m}\sum_{i=1}^{2m}\log\frac{\exp(S_{i,y_i})}{\sum_{j=1}^{2m}\exp(S_{i,j})}, \tag{2}$$

where X denotes the input augmented textual embeddings $\widetilde{\boldsymbol{E}}^t$ or augmented image embeddings $\widetilde{\boldsymbol{E}}^i$. $\exp(S_{i,y_i})$ represents the exponentiated similarity between the current view and its corresponding positive view, while $\exp(S_{i,j})$ denotes the exponentiated similarity between the current view and all

views. Based on this formulation, we obtain the overall intra-modal contrastive learning loss for the textual and image modalities of knowledge concepts as follows:

$$\mathcal{L}_{intra} = \mathcal{L}^t_{intra} + \mathcal{L}^i_{intra}. \tag{3}$$

Following bidirectional cross-modal enhancement, we adopt residual connections to fuse features. Specifically, enhanced textual and image embeddings are added to their corresponding initial CLIP embeddings, followed by ReLU activation to introduce nonlinearity. This design integrates cross-modal information while preserving original textual semantic and image discriminative characteristics. Subsequently, a feed-forward network projects the fused features into a unified vector space to mitigate distributional discrepancies between modalities. The final concept embeddings are formulated as follows:

$$\widehat{\boldsymbol{E}}^t = \mathrm{FFN}(\tilde{\boldsymbol{E}}^t + \boldsymbol{E}^t); \quad \widehat{\boldsymbol{E}}^i = \mathrm{FFN}(\tilde{\boldsymbol{E}}^i + \boldsymbol{E}^i), \tag{4}$$

where $\hat{\boldsymbol{E}}^t, \hat{\boldsymbol{E}}^i \in \mathbb{R}^{m\times 256}$ denote the aligned textual and image representations, respectively.

To promote semantic consistency between modalities, we employ inter-modal contrastive learning optimized via a bidirectional InfoNCE objective with hard negative enhancement. Specifically, we apply ℓ_2 normalization to the representations and construct the similarity matrix as follows:

$$\widehat{\boldsymbol{E}}^{t\prime} = \frac{\widehat{\boldsymbol{E}}^t}{||\widehat{\boldsymbol{E}}^t||_2}; \quad \widehat{\boldsymbol{E}}^{i\prime} = \frac{\widehat{\boldsymbol{E}}^i}{||\widehat{\boldsymbol{E}}^i||_2}; \quad \boldsymbol{S} = \frac{\widehat{\boldsymbol{E}}^{t\prime}(\widehat{\boldsymbol{E}}^{i\prime})^T}{\tau}, \tag{5}$$

where $\boldsymbol{S} \in \mathbb{R}^{m\times m}$ denotes the inter-modal similarity matrix, and τ is the temperature scaling parameter. We then adopt a bidirectional InfoNCE objective to jointly constrain both directions, namely text-to-image retrieval and image-to-text retrieval:

$$\mathcal{L}_{t2i} = -\frac{1}{m}\sum_{k=1}^{m}\log\frac{\exp(\boldsymbol{S}_{k,k})}{\sum_{j=1}^{m}\exp(\boldsymbol{S}_{k,j})}; \quad \mathcal{L}_{i2t} = -\frac{1}{m}\sum_{k=1}^{m}\log\frac{\exp(\boldsymbol{S}_{k,k})}{\sum_{j=1}^{m}\exp(\boldsymbol{S}_{j,k})}, \tag{6}$$

where $\boldsymbol{S}_{k,k}$ denotes the similarity of the matched text–image pair corresponding to the same concept, and $\boldsymbol{S}_{k,j}(j \neq k)$ represents the similarities of negative pairs.

Relying on random sampling often leads to easy negatives dominating training, hindering the discrimination of semantically similar concepts. To address this, we introduce hard negative mining. Specifically, we exclude positive pairs via a diagonal mask and select the highest-similarity samples as hard negatives:

$$\boldsymbol{S}^{\mathrm{mask}} = \boldsymbol{S} \odot (\mathbf{1}_m\mathbf{1}_m^\top - \boldsymbol{I}_m), \quad s_i^{\mathrm{hard}} = \max_{j\neq i}\boldsymbol{S}^{\mathrm{mask}}_{i,j}, \tag{7}$$

where $\odot$ denotes the Hadamard product, $\boldsymbol{I}_m$ is the $m \times m$ identity matrix, and s_i^{hard} denotes the similarity of the most confusing negative sample with respect to the i-th anchor sample.

Based on the hardest negative samples, we formulate a hardest-negative loss that explicitly encourages a larger margin between the similarity of positive pairs and that of the hardest negative pairs:

$$\mathcal{L}_{hard} = -\frac{1}{m}\sum_{i=1}^{m}\log\frac{\exp(\boldsymbol{S}_{i,i})}{\exp(\boldsymbol{S}_{i,i}) + \exp(s_i^{hard})}. \tag{8}$$

Finally, we combine the bidirectional InfoNCE-based alignment objective with the hard negative loss in a weighted manner:

$$\mathcal{L}_{inter} = \frac{1}{2}(\mathcal{L}_{t2i} + \mathcal{L}_{i2t}) + \lambda\mathcal{L}_{hard}, \tag{9}$$

where λ is a weighting coefficient that balances the optimization objectives.

3.4 Concept Context Feature Embedding Module

In concept prerequisite relation learning, learning resources serve as the carriers of concepts. The pedagogical ordering dependencies among learning resources implicitly reflect the prerequisite logic among the concepts they contain. To capture the rich semantic content of learning resources, we first encode the textual content of each resource using a pretrained BERT (Devlin et al., 2019) model. Subsequently, we construct a learning resource dependency graph $G_R = (V, E_R)$ based on the

dependency relationships among learning resources, where V denotes the set of learning resource nodes and E_R represents the set of directed edges between resources. Given the crucial role of edge directionality in information propagation, we separately compute each node's in-degree and out-degree and normalize the adjacency matrices accordingly. Inspired by previous work (Cui et al., 2024), we employ a Directed Graph Convolutional Network (DGCN) (Tong et al., 2020; Zhang et al., 2021) to learn resource embeddings from G_R.

For the learning resource dependency graph, we perform information propagation in parallel along both the in-degree and out-degree directions:

$$\boldsymbol{x}_i^{(l+1),s} = \phi\left(\sum_{v_j \in N(i)^s} \frac{A_{j,i}^{(s)}}{\sqrt{\deg_i^{(s)} \cdot \deg_j^{(-s)}}} \boldsymbol{\Theta}_N^{(l+1),s} \boldsymbol{x}_j^{(l)}\right), \tag{10}$$

where $s \in \{+, -\}$ denotes message propagation along the out-degree and in-degree directions, respectively. $N(i)^s$ represents the neighbor set of node v_i under direction s, including node v_i itself. A and $\deg$ represent normalized edge weights and node degrees. $\boldsymbol{\Theta}_N^{(l+1),s}$ is the learnable weight matrix, and $\boldsymbol{x}_j^{(l)}$ denotes the embedding of node v_j at layer l. $\phi(\cdot)$ is a nonlinear activation function. Initial embeddings $\boldsymbol{x}_j^{(0)} \in \mathbb{R}^{768}$ are from BERT. Subsequently, we fuse the directional embeddings:

$$\boldsymbol{x}_i^{(l+1)} = \phi\left(\mathrm{Linear}(\boldsymbol{x}_i^{(l+1),+}) + \mathrm{Linear}(\boldsymbol{x}_i^{(l+1),-})\right). \tag{11}$$

Finally, we obtain the concept embedding $\boldsymbol{x}_i^r \in \mathbb{R}^{256}$ from the learning resource dependency graph.

After obtaining learning resource embeddings, we construct context features. A gating mechanism enables adaptive fusion across modalities, yielding the multimodal concept embedding $\boldsymbol{E}^H \in \mathbb{R}^{256}$. To integrate multimodal features with contextual information from resources, we employ multi-head cross-attention. Specifically, $\boldsymbol{E}^H$ serves as query, while the mean embedding of K associated resources acts as key and value:

$$\bar{\boldsymbol{X}}_c^r = \frac{1}{K}\sum_{r=1}^{K} \boldsymbol{E}_{(c)}^r; \quad \widehat{\boldsymbol{E}}_c = \mathrm{MultiAtt}(\boldsymbol{E}_c^H; \bar{\boldsymbol{X}}_c^r; \bar{\boldsymbol{X}}_c^r), \tag{12}$$

where $\bar{\boldsymbol{X}}_c^r$ denotes the average embedding of resources associated with concept c, and $\hat{\boldsymbol{E}}_c \in \mathbb{R}^{256}$ represents its context feature.

3.5 Relation Prediction Module

In feature preprocessing, we concatenate multimodal and context features, projecting them to a lower-dimensional space:

$$\boldsymbol{E}_c = \mathrm{Linear}([\boldsymbol{E}_c^H \parallel \widehat{\boldsymbol{E}}_c]), \tag{13}$$

where $\parallel$ denotes concatenation, and $\boldsymbol{E}_c \in \mathbb{R}^{256}$ represents the final representation of concept c.

We employ a Siamese network (Bromley et al., 1993) to predict whether concept c_i is a prerequisite for concept c_j:

$$\hat{y}(c_i, c_j) = \sigma(\boldsymbol{W}_p \cdot [\boldsymbol{E}_i \parallel \boldsymbol{E}_j \parallel (\boldsymbol{E}_i - \boldsymbol{E}_j) \parallel (\boldsymbol{E}_i \odot \boldsymbol{E}_j)] + \boldsymbol{b}_p), \tag{14}$$

where $\boldsymbol{W}_p$ and $\boldsymbol{b}_p$ are the parameter weight matrix and bias vector, respectively. The prediction loss is computed via binary cross-entropy:

$$\mathcal{L}_{preq} = \frac{1}{|D|}\sum_{(c_i,c_j)\in D} \mathrm{BCE}(\hat{y}(c_i, c_j), y_{c_i c_j}), \tag{15}$$

where D denotes the dataset and $y_{c_i c_j} \in \{0, 1\}$ is the ground-truth label. The overall loss function is:

$$\mathcal{L} = \mathcal{L}_{preq} + \alpha\mathcal{L}_{inter} + \beta\mathcal{L}_{intra}, \tag{16}$$

where α and β are balance coefficients. The complete procedure is summarized in Appendix B (Algorithm 1).

4 EXPERIMENTAL SETUP

In this section, we describe the experimental configuration. To comprehensively evaluate MMCPL, our experiments address four questions: (1) How does MMCPL perform compared to baseline methods? (2) What is the impact of various components? (3) Does the cross-modal attention mechanism effectively align text and image modalities? (4) How sensitive is the model to hyperparameter configurations and what are training costs?

4.1 DATASETS

We conduct experiments on three widely adopted public benchmark datasets:

- **University Course Dataset (UCD)**[1]: Covers computer science topics (Liang et al., 2017) sourced from course materials across 11 universities in the US.
- **LectureBank**[2]: Originates from an online education platform (Li et al., 2019) across five domains: NLP, ML, AI, deep learning, and information retrieval.
- **MOOC**[3]: Compiles subtitles from 38 computer science video playlists (Liang et al., 2017).

Table 1: Dataset statistics. $\mathcal{C}$, $\mathcal{R}$, and $\mathcal{P}$ denote concepts, learning resources, and concept pairs, respectively.

Dataset	$\lvert\mathcal{C}\rvert$	$\lvert\mathcal{R}\rvert$	$\lvert\mathcal{P}^+\rvert$	$\lvert\mathcal{P}^-\rvert$
UCD	407	654	4347	4347
LectureBank	246	277	601	600
MOOC	406	381	4332	4332

4.2 BASELINE METHODS

We compare MMCPL with nine existing concept prerequisite prediction methods: RefD (Liang et al., 2015), PREREQ (Roy et al., 2019), R-VGAE(T) and R-VGAE(P) (Li et al., 2020a), MHAVGAE (Zhang et al., 2022), HGAPNet (Mazumder et al., 2023), LCPRE (Sun et al., 2024), GKROM (Zhang et al., 2025b), and LECDPR (Xiao et al., 2026). Detailed descriptions of all baselines are provided in Appendix C.1.

4.3 EVALUATION METRICS AND IMPLEMENTATION DETAILS

Evaluation Metrics. We adopt Accuracy (ACC), F1 score (F1), and Area Under ROC Curve (AUC). All experiments were repeated five times with different random seeds, reporting mean values with paired t-tests against the best baseline.

Implementation Details. Following GKROM, we partitioned datasets into train, validation, and test sets in an 8:1:1 ratio. Models were trained for 40 epochs using Adam with seed 25, learning rate 10^{-4}, weight decay 10^{-3}, and batch size 16. Missing images were synthesized by Wan2.5-t2i-preview. CLIP generated 512-dim text/image embeddings, and BERT provided 768-dim resource vectors. Experiments ran on an RTX 4090 GPU.

5 EXPERIMENTAL RESULTS

5.1 OVERALL PERFORMANCE

Table 2 reports comparative results across the three datasets: (1) MMCPL consistently achieves the best performance across the majority of benchmark datasets and metrics, particularly showing dominant advantages on large-scale datasets (MOOC and UCD), with statistically significant improvements ($p < 0.01$ or $p < 0.05$) over the strongest baseline. (2) Among the compared models, GKROM and HGAPNet achieve the strongest overall results, especially on the MOOC and UCD datasets. (3) The improvements of MMCPL are consistent across datasets and metrics, demonstrating the effectiveness of multimodal representation learning for prerequisite relation prediction.

[1] https://github.com/suderoy/PREREQ-IAAI-19

[2] https://github.com/Yale-LILY/LectureBank

[3] https://github.com/suderoy/PREREQ-IAAI-19

Table 2: Performance comparison. The best result in each column is highlighted in **bold**, and the second-best is underlined.

Model	LectureBank			MOOC			UCD		
	ACC	F1	AUC	ACC	F1	AUC	ACC	F1	AUC
VGAE	0.595	0.713	0.653	0.666	0.665	0.714	0.693	0.692	0.755
RefD	0.585	0.468	0.630	0.587	0.414	0.490	0.762	0.711	0.752
PREREQ	0.512	0.289	0.481	0.542	0.574	0.624	0.543	0.586	0.670
R-VGAE(T)	0.562	0.552	0.542	0.592	0.543	0.660	0.684	0.661	0.764
R-VGAE(P)	0.528	0.400	0.529	0.534	0.410	0.873	0.736	0.722	0.832
MHAVGAE	0.834	0.830	0.901	0.747	0.764	0.875	0.787	0.795	0.864
HGAPNet	0.793	0.793	0.873	0.875	0.878	0.938	0.868	0.871	0.945
LCPRE	0.818	0.810	0.880	0.825	0.822	0.889	0.836	0.821	0.888
GKROM	0.785	0.779	0.862	0.879	0.884	0.941	0.869	0.874	0.939
LECDPR	0.851	0.854	**0.936**	0.890	0.888	0.949	0.861	0.854	0.922
MMCPL(ours)	**0.859**	**0.859**	0.907	**0.894**	**0.903**	**0.950**	**0.901**	**0.905**	**0.950**

5.2 Ablation Experiment

Table 3 shows the ablation study findings: (1) Removing text caused severe degradation, establishing text as the primary foundation; removing images also led to consistent drops, proving the visual modality's complementary value. (2) Removing intra- or inter-modal contrastive learning degraded performance, with inter-modal alignment proving particularly vital for cross-modal fusion. (3) Removing cross-modal attention (CMA) confirmed that simple fusion is insufficient compared to bidirectional attention interactions. Additionally, an in-depth investigation into the impact of the resource graph modeling (DGCN) under different graph topologies is detailed in Appendix C.2.

Table 3: The ablation study of MMCPL.

Method	LectureBank			MOOC			UCD		
	ACC	F1	AUC	ACC	F1	AUC	ACC	F1	AUC
Ours MMCPL	0.859	0.859	0.907	0.894	0.903	0.950	0.901	0.905	0.950
Ours w/o text	0.801	0.800	0.887	0.878	0.881	0.946	0.875	0.876	0.944
Ours w/o image	0.826	0.823	0.886	0.889	0.898	0.949	0.888	0.887	0.941
Ours w/o Intra-CL	0.826	0.820	0.895	0.892	0.900	0.950	0.892	0.896	0.940
Ours w/o Inter-CL	0.818	0.819	0.888	0.882	0.885	0.941	0.890	0.894	0.949
Ours w/o CMA	0.818	0.813	0.892	0.885	0.895	0.947	0.880	0.882	0.946

5.3 Feature Clustering Visualization

Figure 3 visualizes t-SNE distributions for 50 concepts from MOOC. Without our strategy (left), modalities are distant and concept clusters entangled. With our strategy (right), modalities for the same concept cluster tightly while inter-concept boundaries remain clear, confirming effective cross-modal alignment.

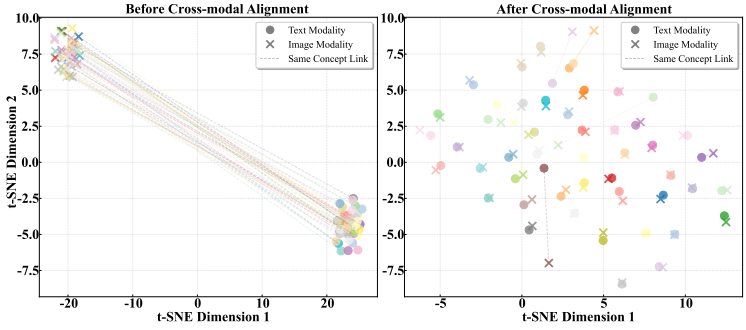

Figure 3: Feature clustering analysis on MOOC.

5.4 Hyperparameter Analysis

As shown in Figure 4: (1) Two directed graph convolution layers ($l = 2$) yields optimal results, avoiding over-smoothing. (2) Gaussian noise intensity around 0.01–0.05 balances view diversity and semantic integrity. (3) Balanced contrastive loss weights of 0.02 maximize consistency and discrimination. (4) Temperature parameters around 0.2 (intra-CL) and 0.05 (inter-CL) deliver the best trade-offs.

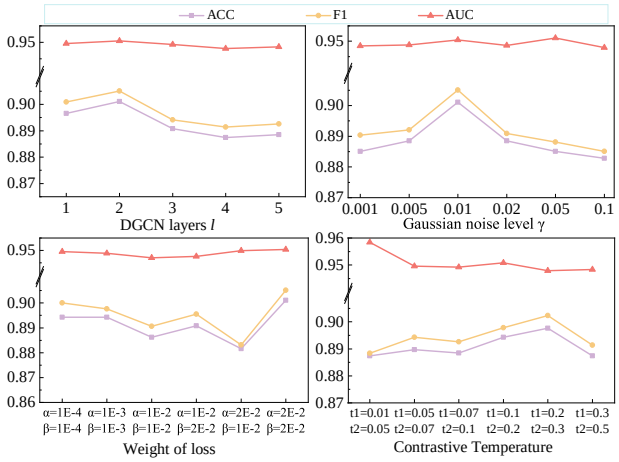

Figure 4: The hyperparameter analysis of MMCPL on UCD.

5.5 Training Overhead Analysis

Figure 5 reports the forward FLOPs per optimizer iteration and the average training time per epoch on MOOC. MMCPL requires 7.68 GFLOPs per iteration, substantially lower than LCPRE (164.03 GFLOPs), while maintaining the shortest training time of 21.09 seconds per epoch among all compared methods. Although GKROM and LECDPR have lower theoretical FLOP counts (0.33 and 0.39 GFLOPs, respectively), their measured training times are considerably longer (83.65 and 214.76 seconds), indicating that sparse or irregular computation can limit practical GPU efficiency. These results demonstrate that MMCPL achieves a favorable balance between computational cost and training efficiency.

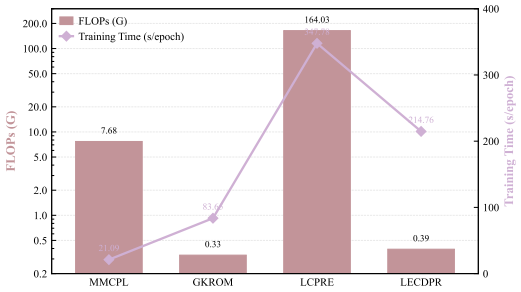

Figure 5: The training overhead analysis on MOOC.

6 Conclusion and Future Work

We presented MMCPL, a multimodal embedding enhancement method for concept prerequisite learning that couples bidirectional cross-modal attention with dual-branch contrastive learning. Extensive experiments confirm state-of-the-art performance across three benchmarks. Future work will investigate domain-specific end-to-end multimodal pretraining, multi-hop transitive prerequisite reasoning, and downstream learning path recommendation.

AI USE STATEMENT

In this work, we used generative AI tools for polishing the manuscript's readability and linguistic quality. Specifically, we utilized a generative language model (GPT-5.4) to refine grammar, polish academic expressions, and enhance sentence clarity in portions of the text. We have not used generative AI tools for core research ideas, technical contributions, method design, experimental setup, analysis, or figure creation, all of which were conceived, developed, and produced solely by the authors. We have reviewed all AI-assisted text and take full responsibility for the final content of this work.

ETHICS STATEMENT

This work focuses on concept prerequisite relation extraction based on public educational datasets and open knowledge bases (Wikipedia). It does not involve human subject experiments, personally identifiable information, or sensitive personal data. There are no anticipated negative societal or ethical implications.

REFERENCES

Jane Bromley, Isabelle Guyon, Yann LeCun, Eduard Säckinger, and Roopak Shah. Signature verification using a siamese time delay neural network. In *NIPS*, pp. 737–744. Morgan Kaufmann, 1993.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey E. Hinton. A simple framework for contrastive learning of visual representations. In *ICML*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1597–1607. PMLR, 2020a.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. UNITER: universal image-text representation learning. In *ECCV (30)*, volume 12375 of *Lecture Notes in Computer Science*, pp. 104–120. Springer, 2020b.

Chaoran Cui, Yumo Yao, Chunyun Zhang, Hebo Ma, Yuling Ma, Zhaochun Ren, Chen Zhang, and James Ko. DGEKT: A dual graph ensemble learning method for knowledge tracing. *ACM Trans. Inf. Syst.*, 42(3):78:1–78:24, 2024.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT (1)*, pp. 4171–4186. Association for Computational Linguistics, 2019.

Yao Dong, Yuxi Liu, Yongfeng Dong, Yacong Wang, and Min Chen. Multi-knowledge enhanced graph convolution for learning resource recommendation. *Knowl. Based Syst.*, 291:111521, 2024.

Weibo Gao, Hao Wang, Qi Liu, Fei Wang, Xin Lin, Linan Yue, Zheng Zhang, Rui Lv, and Shijin Wang. Leveraging transferable knowledge concept graph embedding for cold-start cognitive diagnosis. In *SIGIR*, pp. 983–992. ACM, 2023.

Yuncheng Guo and Xiaodong Gu. MMRL: multi-modal representation learning for vision-language models. In *CVPR*, pp. 25015–25025. Computer Vision Foundation / IEEE, 2025.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A. Shamma, Michael S. Bernstein, and Li Fei-Fei. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *Int. J. Comput. Vis.*, 123(1):32–73, 2017.

Irene Li, Alexander R. Fabbri, Robert R. Tung, and Dragomir R. Radev. What should I learn first: Introducing lecturebank for NLP education and prerequisite chain learning. In *AAAI*, pp. 6674–6681. AAAI Press, 2019.

Irene Li, Alexander R. Fabbri, Swapnil Hingmire, and Dragomir R. Radev. R-VGAE: relational-variational graph autoencoder for unsupervised prerequisite chain learning. In *COLING*, pp. 1147–1157. International Committee on Computational Linguistics, 2020a.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation. In *ICML*, volume 162 of *Proceedings of Machine Learning Research*, pp. 12888–12900. PMLR, 2022.

Xiujun Li, Xi Yin, Chunyuan Li, Pengchuan Zhang, Xiaowei Hu, Lei Zhang, Lijuan Wang, Houdong Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *ECCV (30)*, volume 12375 of *Lecture Notes in Computer Science*, pp. 121–137. Springer, 2020b.

Yi Li, Hualiang Wang, Yiqun Duan, Jiheng Zhang, and Xiaomeng Li. A closer look at the explainability of contrastive language-image pre-training. *Pattern Recognit.*, 162:111409, 2025.

Chen Liang, Zhaohui Wu, Wenyi Huang, and C. Lee Giles. Measuring prerequisite relations among concepts. In *EMNLP*, pp. 1668–1674. The Association for Computational Linguistics, 2015.

Chen Liang, Jianbo Ye, Zhaohui Wu, Bart Pursel, and C. Lee Giles. Recovering concept prerequisite relations from university course dependencies. In *AAAI*, pp. 4786–4791. AAAI Press, 2017.

Weixin Liang, Yuhui Zhang, Yongchan Kwon, Serena Yeung, and James Y. Zou. Mind the gap: Understanding the modality gap in multi-modal contrastive representation learning. In *NeurIPS*, 2022.

Jingping Liu, Mingchuan Zhang, Weichen Li, Chao Wang, Shuang Li, Haiyun Jiang, Sihang Jiang, Yanghua Xiao, and Yunwen Chen. Beyond entities: A large-scale multi-modal knowledge graph with triplet fact grounding. In *AAAI*, pp. 18653–18661. AAAI Press, 2024.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *NeurIPS*, pp. 13–23, 2019.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *NeurIPS*, 2022.

Debjani Mazumder, Jiaul H. Paik, and Anupam Basu. A graph neural network model for concept prerequisite relation extraction. In *CIKM*, pp. 1787–1796. ACM, 2023.

Xianghan (Christine) O'Dea and Julian Stern. Virtually the same?: Online higher education in the post covid-19 era. *Br. J. Educ. Technol.*, 53(3):437–442, 2022.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *ICML*, volume 139 of *Proceedings of Machine Learning Research*, pp. 8748–8763. PMLR, 2021.

Sudeshna Roy, Meghana Madhyastha, Sheril Lawrence, and Vaibhav Rajan. Inferring concept prerequisite relations from online educational resources. In *AAAI*, pp. 9589–9594. AAAI Press, 2019.

Mohsen Sayyadiharikandeh, Jonathan Gordon, José Luis Ambite, and Kristina Lerman. Finding prerequisite relations using the wikipedia clickstream. In *WWW (Companion Volume)*, pp. 1240–1247. ACM, 2019.

Lingyun Song, Mengting He, Xuequn Shang, Chen Yang, Jun Liu, Mengzhen Yu, and Yu Lu. A deep cross-modal neural cognitive diagnosis framework for modeling student performance. *Expert Syst. Appl.*, 230:120675, 2023.

Hao Sun, Yuntao Li, and Yan Zhang. Conlearn: Contextual-knowledge-aware concept prerequisite relation learning with graph neural network. In *SDM*, pp. 118–126. SIAM, 2022.

Jingwen Sun, Yu He, Yiyu Xu, Jingwei Sun, and Guangzhong Sun. A learning-path based supervised method for concept prerequisite relations extraction in educational data. In *CIKM*, pp. 2168–2177. ACM, 2024.

Partha P. Talukdar and William W. Cohen. Crowdsourced comprehension: Predicting prerequisite structure in wikipedia. In *BEA@NAACL-HLT*, pp. 307–315. The Association for Computer Linguistics, 2012.

Hao Tan and Mohit Bansal. LXMERT: learning cross-modality encoder representations from transformers. In *EMNLP/IJCNLP (1)*, pp. 5099–5110. Association for Computational Linguistics, 2019.

Zekun Tong, Yuxuan Liang, Changsheng Sun, David S. Rosenblum, and Andrew Lim. Directed graph convolutional network. *CoRR*, abs/2004.13970, 2020.

Chao Wang, Hengshu Zhu, Peng Wang, Chen Zhu, Xi Zhang, Enhong Chen, and Hui Xiong. Personalized and explainable employee training course recommendations: A bayesian variational approach. *ACM Trans. Inf. Syst.*, 40(4):70:1–70:32, 2022.

Shuting Wang and Lei Liu. Prerequisite concept maps extraction for automaticassessment. In *WWW (Companion Volume)*, pp. 519–521. ACM, 2016.

Zhichun Wang, Yifeng Shao, Boci Peng, Bangui Li, Yun Li, Qianren Wang, and Nijun Li. Multi-view transformer-based network for prerequisite learning in concept graphs. In *ISWC (1)*, volume 15231 of *Lecture Notes in Computer Science*, pp. 67–86. Springer, 2024.

Kui Xiao, Lele Zheng, Xiaoxue He, Miao Zhang, Zhifang Huang, and Yan Zhang. LECDPR: LLM enhancement and concept-document interactive modeling for prerequisite relation prediction. In *IJCAI*, pp. 5092–5100. ijcai.org, 2026.

Shenbao Yu, Yifeng Zeng, Fan Yang, and Yinghui Pan. Causal-driven skill prerequisite structure discovery. In *AAAI*, pp. 20604–20612. AAAI Press, 2024.

Juntao Zhang, Nanzhou Lin, Xuelong Zhang, Wei Song, Xiandi Yang, and Zhiyong Peng. Learning concept prerequisite relations from educational data via multi-head attention variational graph auto-encoders. In *WSDM*, pp. 1377–1385. ACM, 2022.

Miao Zhang, Jiawei Wang, Kui Xiao, Zhifang Huang, Zhifei Li, and Yan Zhang. Enhancing weak supervision for concept prerequisite relation learning. *IEEE Trans. Big Data*, 11(5):2643–2656, 2025a.

Miao Zhang, Jiawei Wang, Kui Xiao, Shihui Wang, Yan Zhang, Hao Chen, and Zhifei Li. Learning concept prerequisite relation via global knowledge relation optimization. In *AAAI*, pp. 1638–1646. AAAI Press, 2025b.

Xitong Zhang, Yixuan He, Nathan Brugnone, Michael Perlmutter, and Matthew J. Hirn. Magnet: A neural network for directed graphs. In *NeurIPS*, pp. 27003–27015, 2021.

Xiangru Zhu, Zhixu Li, Xiaodan Wang, Xueyao Jiang, Penglei Sun, Xuwu Wang, Yanghua Xiao, and Nicholas Jing Yuan. Multi-modal knowledge graph construction and application: A survey. *IEEE Trans. Knowl. Data Eng.*, 36(2):715–735, 2024.

A Concept Multimodal Representation Generation and Evaluation

A.1 Concept Multimodal Representation Generation

To acquire multimodal information, we retrieve Wikipedia pages based on concept names, extracting the first 400 words as textual modality and the first available image as the image modality. Since some concept entries lack images, we employ a Large Language Model to construct semantic prompts from the corresponding concept descriptions and use Wan2.5-t2i-preview to synthesize the missing visual content following the prompt design shown in Figure 6. The missing-image ratios for UCD, LectureBank, and MOOC are 42.26%, 43.50%, and 40.64%, respectively. Generated images are subsequently filtered for semantic consistency with the concept description and visual quality; only retained images are used for multimodal representation learning.

System prompts:
You are an educational illustrator. Based on the following concept, create a minimalist, clear, easy-to-understand teaching illustration. The illustration should be monochrome or near-monochrome, flat, and structurally clear with simple geometric shapes.

Input prompts:
[Concept Name]: {concept_name};

[Concept Description]: {concept_context};

Illustration goals:
Represent the core idea using the simplest visual form;
Understandable at a glance;
.....

Output:
Images generated through text-to-image model.

Figure 6: Prompt design for image modality completion.

After obtaining the textual and image modalities of each concept, we employ the pretrained CLIP model to encode the two modalities separately:

$$\boldsymbol{E}^t = \mathrm{CLIP}(T); \quad \boldsymbol{E}^i = \mathrm{CLIP}(V), \tag{17}$$

where $\boldsymbol{E}^t, \boldsymbol{E}^i \in \mathbb{R}^{m\times 512}$. T and V denote the textual and image modalities of the concepts, respectively, and m is the number of concepts.

A.2 Performance on Real and Synthetic Image Subsets

Using the concept totals reported in Table 1, UCD contains 235 real-image and 172 synthetic-image concepts; LectureBank contains 139 real-image and 107 synthetic-image concepts; and MOOC contains 241 real-image and 165 synthetic-image concepts. The corresponding synthesis ratios are 42.26%, 43.50%, and 40.64%, respectively.

To evaluate the quality of the synthesized visual modality and its impact on prerequisite relation learning, we partitioned the test set into a **REAL** subset (where concept pairs have original images from authoritative sources) and a **SYNTHETIC** subset (where at least one concept relies on generated images). As shown in Table 4, MMCPL achieves superior performance across all datasets. Specifically, on LectureBank and UCD, the REAL subset exhibits the highest overall accuracy and AUC (e.g., 0.9770 AUC on LectureBank and 0.9603 AUC on UCD), verifying that original high-quality visual representations provide rich and clean structural cues for dependency inference. Meanwhile, the SYNTHETIC subset maintains comparably strong performance, achieving over 0.84 ACC and 0.87 AUC across all benchmarks. These observations demonstrate that our cross-modal enhanced embedding module and dual-branch contrastive objectives effectively mitigate generation noise, proving that LLM-driven visual completion reliably resolves the missing modality dilemma in educational knowledge graphs.

Table 4: Performance comparison of MMCPL on real image subset (REAL) and synthesized image subset (SYNTHETIC) across three benchmark datasets.

Dataset	Data Split	ACC	F1	AUC
LectureBank	REAL	0.8940	0.8820	0.9770
	SYNTHETIC	0.8430	0.8500	0.8740
MOOC	REAL	0.8871	0.9103	0.9448
	SYNTHETIC	0.9006	0.8967	0.9535
UCD	REAL	0.9249	0.9381	0.9603
	SYNTHETIC	0.8855	0.8765	0.9413

B Algorithm Details

Algorithm 1: The overall process of MMCPL.

Input: Text embeddings $\mathbf{T} \in \mathbb{R}^{N \times d}$, image embeddings $\mathbf{I} \in \mathbb{R}^{N \times d}$, resource embeddings $\mathbf{R} \in \mathbb{R}^{M \times d'}$, directed adjacency matrices $\mathbf{A}_{\text{out}}, \mathbf{A}_{\text{in}} \in \mathbb{R}^{M \times M}$, concept-resource mapping Φ, concept pair (c_1, c_2). The learnable parameters Θ.
Output: The parameters of MMCPL Θ.
for *each concept* c_k **do**
  Cross-modal attention enhancement:
  Enhance text and image embeddings via cross-modal attention by equation 1.
  Compute intra-modal contrastive loss $\mathcal{L}_{intra}$ by equation 2.
  Modality projection and gated fusion:
  Project to shared space and compute inter-modal loss $\mathcal{L}_{inter}$.
  Fuse text and image representations using a learnable gating mechanism.
  Directed graph convolution for resources:
  Apply directed GCN on $\mathbf{R}$ with $\mathbf{A}_{out}$ and $\mathbf{A}_{in}$ by equation 10 and equation 11.
  Combine directional outputs to obtain resource representations $\mathbf{X}^r$.
end
Concept-resource cross-attention and prediction:
for *each concept* c_k *in* (c_1, c_2) **do**
  Compute the context feature using concept-resource cross-attention by equation 12.
  Compute prerequisite probability using the Siamese network by equation 14.
end
Compute total loss: $\mathcal{L} = \mathcal{L}_{preq} + \alpha\mathcal{L}_{inter} + \beta\mathcal{L}_{intra}$.
return Θ.

C Experimental Details and Additional Analysis

C.1 Baseline Details

- **RefD (Liang et al., 2015):** Uses TF-IDF and a citation-based distance metric.
- **PREREQ (Roy et al., 2019):** Learns representations via Pairwise-LDA and predicts relations with a Siamese network.
- **R-VGAE(T) and R-VGAE(P) (Li et al., 2020a):** Integrate R-GCN with VGAE using term frequencies and phrase vectors, respectively.
- **MHAVGAE (Zhang et al., 2022):** End-to-end framework combining multi-head attention with VGAE.
- **HGAPNet (Mazumder et al., 2023):** Heterogeneous graph neural network with node-level attention.
- **LCPRE (Sun et al., 2024):** Constructs concept learning paths to capture interactions along directed paths.
- **GKROM (Zhang et al., 2025b):** Jointly models multiple knowledge relations via multi-objective optimization.

- **LECDPR (Xiao et al., 2026):** Combines LLM enhancement with concept-document interactive modeling for prerequisite relation prediction.

C.2 Impact of Resource Dependency Modeling (DGCN)

To evaluate the contribution of topological resource dependencies, we compare the complete MMCPL against a variant without the Directed Graph Convolutional Network (`MMCPL w/o DGCN`), where resource context representations are aggregated purely via unweighted pooling without directional message passing.

As shown in Table 5, incorporating DGCN leads to substantial performance gains on the **MOOC** and **UCD** datasets (e.g., F1 improves from 0.850 to 0.903 on MOOC, and from 0.851 to 0.905 on UCD). In these benchmarks, curriculum resources are richly interconnected with well-defined pedagogical sequences, allowing directed graph convolutions to effectively capture high-order dependencies and information propagation paths.

Conversely, on **LectureBank**, removing DGCN yields competitive accuracy (0.859 for both) and slightly higher F1 (0.866 vs. 0.859) and AUC (0.927 vs. 0.907). We attribute this divergence and our architectural choices to the following intrinsic factors:

1. **Topological Sparsity and Structural Noise:** As detailed in Table 1, LectureBank comprises merely 277 learning resources and 601 prerequisite pairs—roughly one-seventh the size of MOOC (4,332 pairs) and UCD (4,347 pairs). On such an overly sparse and fragmented graph, two-layer directional convolutions inevitably force nodes to aggregate distant or weakly related resource context, diluting discriminative representations with structural noise.
2. **Modality Dominance in Concentrated Domains:** In specialized, domain-concentrated corpora like LectureBank, intrinsic semantic and visual representations already exhibit strong discriminative fidelity (as evidenced by the 0.9770 AUC on its clean real-image subset in Table 4). Under high-fidelity multimodal representations, forcing message passing over sparse, noisy topologies risks perturbing the well-aligned feature manifolds.
3. **Behavior of Cross-Attention and Architectural Generality:** In Section 3.4, MMCPL employs a multi-head cross-attention mechanism to adaptively integrate resource features. However, under end-to-end optimization with limited supervisory signals (only $\sim$480 training pairs on LectureBank), cross-attention weights cannot fully degrade to zero without risking severe overfitting. Rather than introducing benchmark-specific heuristic thresholding or ad-hoc gating rules that compromise model generality, we maintain a unified architecture ($l = 2$ layers across all datasets). In practical deployment, `MMCPL w/o DGCN` serves as a lightweight inductive prior for sparse corpora, while the full MMCPL is strongly recommended for interconnected, realistic educational networks. Developing density-aware adaptive graph gating remains a promising direction for future work.

Table 5: Performance comparison with and without DGCN across three benchmark datasets.

Method	LectureBank			MOOC			UCD		
	ACC	F1	AUC	ACC	F1	AUC	ACC	F1	AUC
Ours MMCPL	0.859	0.859	0.907	0.894	0.903	0.950	0.901	0.905	0.950
Ours w/o DGCN	0.859	0.866	0.927	0.854	0.850	0.937	0.851	0.851	0.936