SciMKG: A Multimodal Knowledge Graph for Science Education with Text, Image, Video and Audio

Tong Lu[1], **Zhichun Wang**[*1,2], **Yaoyu Zhou**[1], **Yiming Guan**[1], **Zhiyong Bai**[3], **Junsheng Du**[4]

[1] School of Artificial Intelligence, Beijing Normal University, Beijing, China
[2] Engineering Research Center of Intelligent Technology and Educational Application, Ministry of Education, Beijing, China
[3] Faculty of Education, Beijing Normal University, Beijing, China
[4] School of intelligent system engineering, Sun Yat-sen University, Guangzhou, China
ethanlu@mail.bnu.edu.cn, zcwang@bnu.edu.cn

Abstract

Knowledge graphs (KGs) play a vital role in intelligent education by offering structured representations of educational content. However, constructing multimodal educational knowledge graphs (EKGs) from diverse open educational resources remains a challenge due to the reliance on costly manual annotations and the lack of multimodal integration. In this work, we propose an automated framework that harnesses the reasoning capabilities of large language models (LLMs) to construct multimodal EKGs from open courses efficiently. In our framework, an Extraction-Verification-Integration-Augmentation pipeline is designed to incrementally extract and refine disciplinary concepts from learning resources. Texts, images, videos and audios are aligned with their corresponding concepts. To ensure semantic consistency across modalities, we propose a cross-modal alignment method based on shared structural and semantic features. Using our framework, we build SciMKG, a large-scale multimodal EKG for Chinese K12 education in sciences (biology, physics, and chemistry), encompassing 1,356 knowledge points, 34,630 multimodal concepts, and 403,400 relational triples. Experimental results show that our method improves concept extraction F1 score by 9% over state-of-the-art baselines; both automatic and human evaluations confirm the robustness of our multimodal alignment method. SciMKG and our construction toolkit will be publicly released to support further research and applications in AI-driven education.

Code — https://github.com/kg-bnu/SciMKG

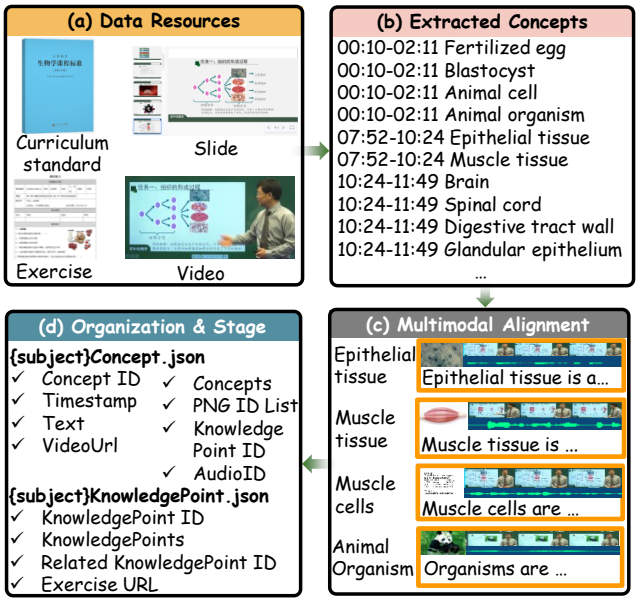

Figure 1: Construction pipeline of SciMKG. Multimodal data (text, image, video, and audio) are processed by the proposed pipeline to identify disciplinary concepts, which are then aligned across modalities using shared structural and semantic features, and indexed for systematic organization.

Introduction

Education is the foundation upon which we build our future. With the advent of the digital era, open educational resources have boomed swiftly, such as billion-scale Massive Open Online Courses (MOOCs) (Billsberry and Alony 2024), Wikipedia knowledge repositories (Moás and Lopes 2023), and ebooks (Karakoç Öztürk 2021) constitute massive multimodal knowledge resources. Knowledge graphs (KGs) have emerged as a powerful paradigm for systematically organizing heterogeneous knowledge. In the education domain, Educational Knowledge Graphs (EKGs) have shown great promise in enabling applications such as such as adaptive question generation (Agrawal et al. 2024), contextual question answering (Wang et al. 2024), and knowledge retrieval (Ma et al. 2024). Most existing efforts focus on extracting textual concepts from curriculum standards, textbooks, and MOOCs (Li et al. 2025). However, the existing EKGs still suffer from two critical limitations in the face of growing educational demands:

- **Reliance on Data Annotation and Model Training.** Many existing EKGs rely heavily on manual annotated training data and hand-crafted deep learning models. For instance, KnowEdu (Chen et al. 2018) trained Named Entity Recognition (NER) models to extract educational concepts, while MOOCCube (Yu et al.

*Corresponding author.

Dataset	Year	Source	Type	Domain	Auto	Modality				Hier	Tool
						T	I	V	A		
LectureBank (Li et al. 2019)	2019	MOOCs	EDR	Edu		✓					
MOOCcube (Yu et al. 2020)	2020	MOOCs	EDR	Edu		✓					
MOOCubeX (Yu et al. 2021)	2021	MOOCs	EDR	Edu		✓					✓
MoocRadar (Yu et al. 2023)	2023	MOOCs	EDR	Edu		✓					✓
KnowEdu (Chen et al. 2018)	2018	CS+Textbook	EKG	Edu		✓				✓	
MEduKG (Li et al. 2022b)	2022	Textbook	EKG	Edu		✓			✓		
SAC-KG (Chen et al. 2024)	2024	SearchEngine	LLMs KG	Gen	✓	✓					
NLP-AKG (Lan et al. 2024)	2024	SearchEngine	LLMs KG	Gen	✓	✓					
GPTEKG (Jhajj et al. 2024a)	2024	MOOCs	LLMs EKG	Edu	✓	✓					
LLM4EduKG (Sun and Zhang 2024)	2024	Textbook	LLMs EKG	Edu	✓	✓					
TIVA-KG (Wang et al. 2023b)	2023	CN+Wiki	MKG	Gen		✓	✓	✓	✓		✓
Ukonw (Gong et al. 2024)	2024	News+Wiki	MKG	Gen		✓	✓				
SciMKG (ours)	2025	MOOCs+CS+ CN+Wiki	LLMs MEKG	Edu	✓	✓	✓	✓	✓	✓	✓

Table 1: Comparison between SciMKG and existing Educational Data Repositories and Knowledge Graphs. **Legend:** Edu = Educational, Gen = General, CS = Curriculum Standards, CN = ConceptNet, Wiki = Wikipedia, EDR = Educational Data Repository, EKG = Educational Knowledge Graph, MKG = Multimodal Knowledge Graph, Hier = Hierarchization, T/I/V/A = Text/Image/Video/Audio, Tool = Toolkit, Auto = Automatic Construction.

2020, 2021) and ACE (Aytekin, Saygın et al. 2024) fine-tuned BERT-based models. These approaches are expensive, lack adaptability, and introduce subjective bias—making them unsuitable for the frequently updated open-domain educational resources. Recently, LLMs have been explored for their zero-shot reasoning capabilities, yet most existing LLM-based approaches, such as LLM4EduKG (Jhajj et al. 2024b), still rely on partially handcrafted KG skeletons.

- **Incomplete Modality.** Recently, there has been a significant amount of work on multimodal applications, such as Multimodal Educational Question Generation (MEQG) (Wu et al. 2024), Multimodal Educational Question Answering (MEQA) (Jin et al. 2024), Cross-Modal Knowledge Retrieval (CKR) (Lerner, Ferret, and Guinaudeau 2024) and Education Agent (EA) (Neira-Maldonado et al. 2024). However, most EKGs are predominantly text-based, limiting their ability to support multimodal educational applications effectively. Although recent works (Bin et al. 2025; Yu et al. 2023, 2021) extract disciplinary knowledge from multimodal heterogeneous resources, the multimodal knowledge is converted into text-based knowledge to simplify the construction of EKGs. TIVA-KG (Wang et al. 2023b) pioneers the inclusion of four modalities of entity attributes, but it relies on manual annotations for multimodal alignment and is limited to general-domain scenarios. With the exponential increase in multi-modal data and the multimodal nature of information flow in real-world educational applications, EKGs are urgent to meet multimodal educational demands.

To address these challenges, we propose an automatic framework for constructing multimodal EKGs and introduce SciMKG, a large-scale, four-modality EKG covering the science disciplines (biology, physics, and chemistry) in K12 education. To the best of our knowledge, SciMKG is the first multimodal EKG covering text, image, video and audio simultaneously (as shown in Table 1). As illustrated in Figure 1, at the heart of our framework is a novel Extraction-Verification-Integration-Augmentation pipeline. It extracts disciplinary concepts from multimodal educational resources using a stepwise reasoning strategy with multiple LLMs to improve accuracy and semantic diversity. The extracted concepts are then semantically aligned across modalities through structural and contextual features. To mitigate the coverage limitations of MOOCs, SciMKG further augments the concept set with knowledge from ConceptNet and Wikipedia. Finally, the graph is structured based on official curriculum standards and systematically organized for downstream educational applications.

Main contributions of this work include:

- We propose an automatic framework for constructing multimodal educational knowledge graphs. We design LLM-powered methods for concept extraction and multimodal alignment.
- We build SciMKG, a multimodal knowledge graph for science education covering text, image, video and audio. It covers 1,356 knowledge points, 34,630 multimodal concepts, and 403,400 relational triples.
- The evaluation demonstrates that our framework outperforms state-of-the-art methods in concept extraction, validates the robustness of multimodal alignment through both automatic and human assessments.

The rest of this paper is organized as follows: Section 2 reviews related work on educational knowledge graphs and multimodal alignment. Section 3 describes the proposed framework for constructing multimodal EKGs, including the concept extraction and multimodal alignment methods, and the built SciMKG. Section 4 presents experimental results

and evaluation. Finally, Section 5 concludes the paper.

Related Work

Conception Extraction. The key stage of EKGs construction is to extract concepts in multi-source heterogeneous educational resources and link them to corresponding attributes. Early research efforts utilized deep learning models to gain semantic and structural features of concepts with manual annotation. The work in (Pan et al. 2017) employed word embedding and graph propagation algorithms to extract and filter candidate concepts, KnowEdu (Chen et al. 2018) utilized a Gated Recurrent Unit network to extract sentence-level concepts. To alleviate the difficulty in manual annotation with the scale of educational data continues to expand, MOOCCubeX (Yu et al. 2021) decomposed concept extraction into three stages and proposed an improved RoBERTa (Cui et al. 2021) model to rank candidate concepts. To further tackle noisy annotation and incomplete annotation, DS-MOCE(Lu et al. 2023) proposed a discipline-aware dictionary and enhanced the quality of concept extraction by distant supervision. LLM4EduKG(Jhajj et al. 2024b) utilized LLMs to automatically extract entities from text without manual annotation. However, these efforts lacked instruction tuning and error correction, leaving the potential of LLMs reasoning capabilities. It is crucial to propose a novel method that guides LLMs to further enhance the accuracy of automatic concept extraction.

Multimodal Alignment. Multimodal alignment (Wallace et al. 2024; Ma et al. 2023) has two categories: explicit alignment and implicit alignment. Explicit alignment employs both supervised and unsupervised models to directly utilize the characteristics of multimodal data for alignment. For instance, Diffusion-DPO (Wallace et al. 2024) leveraged unsupervised diffusion model and AltO (Song et al. 2024) trained a registration network to align multimodal data. Implicit alignment maps different modal data into a shared latent space by utilizing attention mechanisms and common semantic information across different modalities. Att-Sinkhorn (Ma et al. 2023) improved the accuracy of multimodal feature alignment by utilizing attention mechanisms to learn the optimal transport problem of probability distributions across different modalities. Although the aforementioned methods have yielded significant results in specific domains, it is necessary to propose an intuitive and automatic multimodal alignment method for educational resources to align multimodal concepts, due to the complexity of deep learning models and the fact that these models are often difficult to deploy and fine-tune in practical applications.

Methodology

We propose an automatic framework for constructing multimodal EKGs from open educational resources, as illustrated in Figure 2. The framework consists of four main components: (1) data acquisition and preprocessing, (2) concept extraction, (3) multimodal alignment, and (4) knowledge organization and storage. This framework integrates text, image, video, and audio data from diverse sources, systematically extracting and aligning educational concepts across modalities to build a unified and structured knowledge graph.

Data Acquisition and Preprocessing

Educational data are sourced from a series of MOOCs platforms[1] that offer a wide range of knowledge in basic education, including videos, slides and exercises. Considering that natural science disciplines (biology, physics and chemistry) are inherently richer in multimodal characteristics compared to other subjects, we obtain basic educational resources for these disciplines. The subtitles of videos are extracted as text to mine concepts, while the images are extracted from slides for enriching multimodal attributes of concepts. For the purposes of ensuring knowledge clarity and enhancing the accessibility of original resources, we leverage Multimodal Large Language Models (MLLMs) to rewrite extracted text, then generate corresponding audio, and record the URL of each original video.

Concept Extraction

The key to constructing multimodal EKGs is to extract concepts from educational resources. Given a text paragraph set $P = \sum_{i=1}^{n} p_i$ extracted from video subtitles, the task of concept extraction aims to obtain a concept set $C = \sum_{j=1}^{m} c_j$ related to a specified discipline d.

The work in (Zhou et al. 2023) demonstrates that LLMs have achieved excellent results in named entities recognition on social media contexts due to powerful reasoning ability. To mitigate the bias introduced by a single model and further enhance the potential ability of LLMs, we leverage multiple LLMs and a stepwise reasoning strategy to obtain sufficient semantic diversity for concept extraction. More specifically, we design an LLM-powered Extraction-Verification-Integration-Augmentation pipeline for concept extraction, as illustrated in Figure 2(b).

- *Extraction.* We encourage LLMs to understand the syntactic structure of each paragraph in P by guiding them to first identify all phrases based on syntactic tags and then extract candidate disciplinary concepts from the recognized nouns and noun phrases.
- *Verification.* To enhance the validation of extraction, each LLM undergoes iterative SELF-REFINE (Madaan et al. 2023) to prune concepts with disciplinary ambiguity, ultimately yielding a refined set of candidate concepts. During the feedback phase, LLMs [2] are encouraged to verify by themselves whether each candidate concept belongs to the given discipline d and to provide a brief explanation for each concept that is potentially pruned. In the subsequent refinement phase, the LLMs perform precise filtering of candidate concepts based on iteratively accumulated feedback.
- *Integration.* We tailor the Self-Consistency (SC) (Wang et al. 2022), a voting-based approach, to merge extracted

[1]The MOOCs are sourced from open course platforms.

[2]The LLMs used include GPT-4o, Gemini-1.5-flash, and DeepSeek-V3, accessed via their official APIs between January and April 2025.

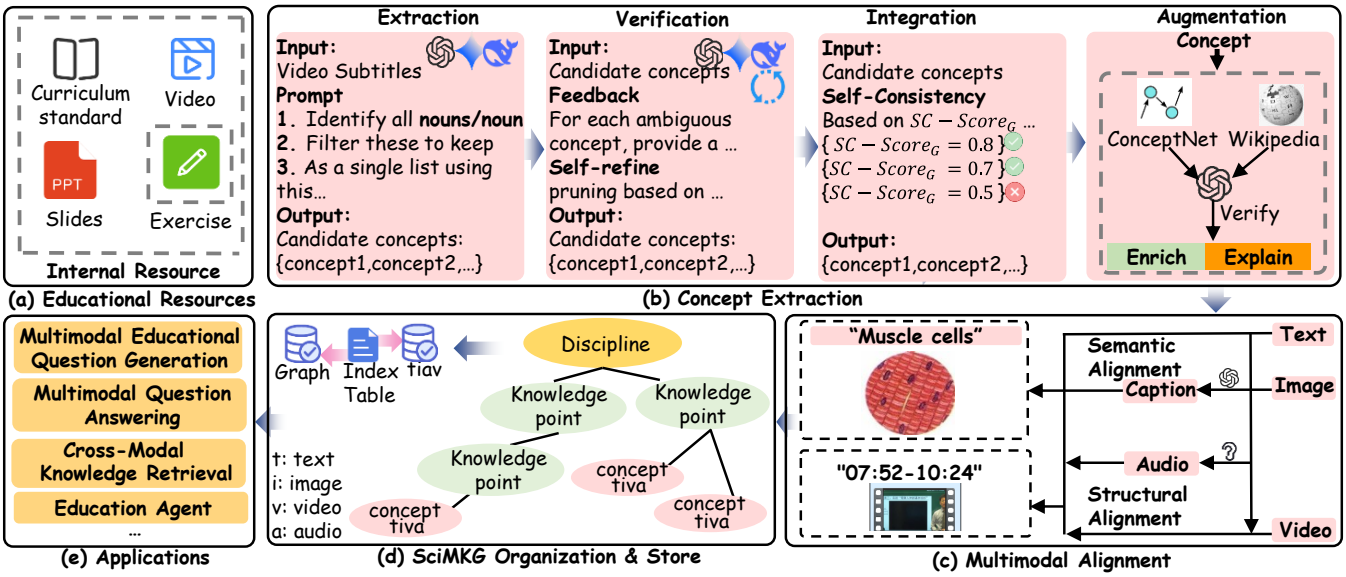

Figure 2: Construction framework of SciMKG.

candidate concepts by multiple LLMs, thereby obtaining the final concept set C. Specifically, each iteration of the SELF-REFINE increments the SC-Score of a concept c_i with the number of iterations $iter$ is set to 5. The confidence threshold α represents the frequency of a candidate concept being identified in validation stage. To account for semantic and knowledge discrepancies caused by different training corpora across LLMs, we define measures:

$$SC{-}Score_G(c_i) = \sum SC{-}Score(c_i, LLM_i), \quad (1)$$

The SC-Score of a candidate concept satisfies $SC{-}Score_G(c_i) \geq \alpha$, which indicates that the candidate concept is validated by all LLMs and included in the final concept set. Otherwise, it is pruned.

- *Augmentation.* In the end, SciMKG is linked to ConceptNet and Wikipedia for expanding disciplinary concepts and obtaining textual explanations, which are subsequently converted into audio. Specifically, each extracted concept c_j in C serves as an initial entry point to interface with ConceptNet. Eight semantic relationships (/r/IsA, /r/RelatedTo, /r/HasContext, /r/PartOf, /r/DefinedAs, /r/Synonym, /r/DerivedFrom and /r/HasPrerequisite) are chosen to identify preliminary candidate expanded concepts using designed regular expression rules. To verify their alignment with the corresponding discipline, LLMs are employed to filter and cross-validate the expanded concepts. Then, refined concepts are linked to Wikipedia to retrieve detailed descriptions, which are summarized using LLMs to produce concise explanations.

Multimodal Alignment

Multimodal alignment is a critical component of our framework, ensuring that concepts extracted from different modalities—text, image, video, and audio—are semantically consistent and interconnected within the knowledge graph. Given a concept set C and multimodal resources $A = \{text, images, videos, audios\}$, the task of multimodal alignment aims to involve mapping the various modal attributes in A to the corresponding concepts in C.

To achieve this, we leverage the capabilities of Multimodal Large Language Models (MLLMs) to understand and align the semantic content across modalities. Specifically, we perform the following steps:

Concept-Image Alignment. We leverage the sophisticated cross-modal understanding of MLLMs [3] to bridge the semantic gap between visual data and concepts. Specifically, an MLLM maps each $image$ in A to a corresponding concept by generating semantic descriptions, while another validates these alignments by assessing their confidence based on $text$ in A. To ensure semantic reliability, only those alignments with confidence scores exceeding the threshold of 0.8 are considered valid.

Concept-Video Alignment. Based on the semantic similarity between a video segment and its co-occurring concepts, we align each video segment in A with concepts in C by jointly leveraging the semantic features of the concept text and the temporal structure from their associated timestamps. Additionally, URLs are provided to ensure convenient access to the original videos.

Concept-Audio Alignment. In light of hallucination challenges with LLMs, the textual explanations $text$ of concepts are rewritten based on knowledge from Wikipedia. The audio modality of concept is generated by TTS based on rewritten explanations.

[3]The MLLMs include GPT-4o and Gemini-1.5-pro during January and April 2025 with official API.

Knowledge Organization and Storage

Knowledge Organization. SciMKG is structured into three levels: discipline, knowledge point and concept. The discipline level represents the uppermost hierarchy, involving disciplines biology, physics and chemistry. The knowledge point level, situated below the discipline level, organizes knowledge according to curriculum standards and the syllabus of MOOCs. Finally, the concept level represents the foundational layer, covering all multimodal concepts.

Knowledge Storage. We propose a new symbolic approach for storing the SciMKG, which ensures the connectivity of knowledge while supporting diverse applications, including knowledge retrieval. Specifically, disciplines, knowledge points, exercises and multimodal concepts within SciMKG are assigned unique identifiers and systematically organized into a hierarchical knowledge graph through cross-indexing mechanisms. As illustrated in Figure 1(d), knowledge points, exercises, concepts and associated multimodal data are interconnected utilizing JSON-formatted cross-reference tables. For enhanced scalability and reusability, SciMKG is also available in an IRI-compliant RDF format.

Statistics and Visualizations

SciMKG includes natural science disciplines for secondary school education covering text, image, video and audio, with a total of 1,356 knowledge points, 34,630 concepts, 403,400 triples, 763 exercises, 10,527 images, 10,425 videos and 34,630 audios. Moreover, SciMKG includes eleven types of relations that connect concepts with multimodal data, which involves the relations "isExplanationOf", "hasExplanation", "isImageOf", "hasImage", "isVideoOf", "hasVideo", "isAudioOf", "hasAudio", and "related to". Each concept and multimodal data are treated as a single entity to support cross-modal knowledge retrieval. Additionally, multimodal concepts are hierarchically connected to knowledge points derived from syllabi, curriculum standards, and disciplines through the relation type "related to". The relations between exercises and knowledge points are defined as "isExerciseOf" and "hasExercise".

The data analysis presents the number of knowledge points, concepts, exercises and triples for each discipline through Table 2, details the statistics of each modality data and the concepts covered by each modality in Table 3, and further illustrates the symbolic hierarchical structure of SciMKG via Figure 4. The Figure 3 demonstrates the percentage of each relation type in SciMKG.

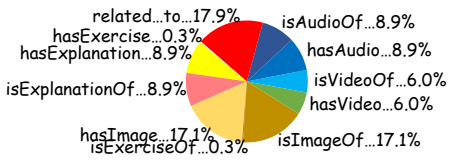

Figure 3: Percentage of each relation type in SciMKG.

Discipline	# Know	# Con	# Exer	# Tri
Biology	526	16,839	255	191,928
Physics	521	11,015	288	145,666
Chemistry	309	6,776	220	65,806

Table 2: Statistics of knowledge in each discipline. Know = knowledge points, Con = concepts, Exer = exercises, Tri = triples.

Modality	# Modality data	% Concepts covered
Image	10,527	0.39
Video	10,425	0.80
Audio	34,630	1.00

Table 3: Statistics of each modality data and the concepts covered by each modality.

Evaluation

In this section, we present a comprehensive evaluation of our framework, assessing its performance from multiple angles, including:

- Evaluation on Concept Extraction: compare our framework with existing SOTA methods; analyze the impact of different LLMs and parameters on the performance; conduct error analysis to identify the limitations of the proposed concept extraction method.
- Evaluation on Modality Alignment: evaluate the effectiveness of modality alignment from both automatic and human evaluation perspectives.

Evaluation Setup

Datasets. A subgraph is randomly selected from the SciMKG dataset to serve as the evaluation dataset. To be specific, the subgraph covers knowledge points from 100 lessons across the disciplines of biology, physics and chemistry, including 4,479 concepts and their corresponding four data modalities.

Models. We employ GPT-4o for direct concept extraction, validate our approach with Gemini 1.5 Flash and DeepSeek-V3, and ensure a fair comparison by also using GPT-4o as the backbone for all LLM-based baselines.

Metrics. We evaluate the concept extraction performance using precision, recall, F1-score, accuracy (ACC), Matthews correlation coefficient (MCC) and area under the curve (AUC), and assess the modality alignment performance using CLIP Score (Radford et al. 2021), BLIP ITM Score (Li et al. 2022a), X-VLM ITM Score (Zeng, Zhang, and Li 2022), VQA Score (Lin et al. 2025) and human evaluation.

System. All experiments are conducted on a server running Ubuntu 22.04 LTS, equipped with an Intel Xeon E5-2686 v4 CPU, 128GB of RAM, and a single NVIDIA GeForce RTX 3090 GPU. The software environment is based on Python 3.7, PyTorch 2.4.0+cu124, and CUDA 12.8.

Hyperparameters. All hyperparameters of the LLM are listed in the Table 6, along with the number of validation

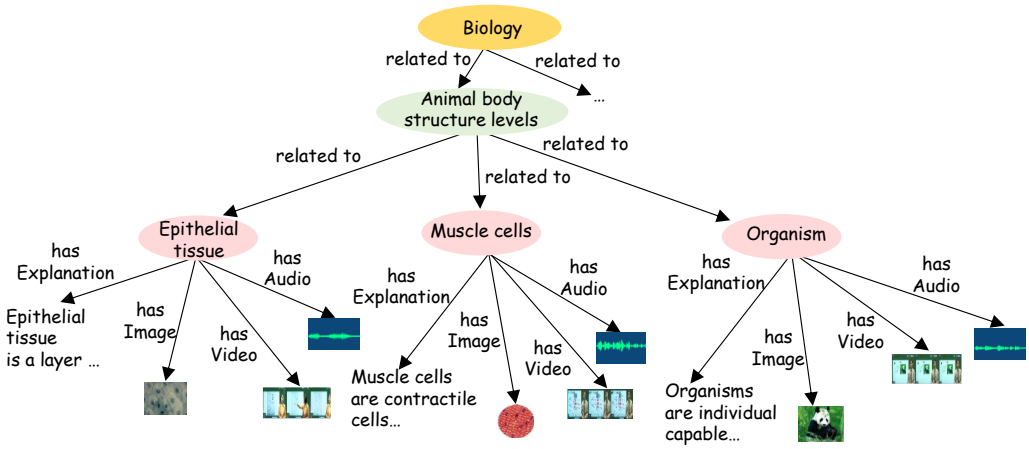

Figure 4: Visualization of a subgraph from SciMKG.

Methods	Precision	Recall	F1-score	ACC	MCC	AUC
Decomposed-QA (Xie et al. 2023)	0.935	0.570	0.712	0.977	**0.741**	0.885
GPT-NER (Wang et al. 2023a)	0.530	**0.931**	0.681	0.923	0.621	0.938
LinkNER (Zhang et al. 2024)	0.855	0.644	0.734	0.981	0.732	0.912
UniversalNER (Zhou et al. 2023)	0.873	0.617	0.718	0.974	0.722	0.929
Self-Improving (Xie et al. 2024)	0.911	0.532	0.670	0.972	0.725	0.935
Our method$_{\text{Q4B+G4B+P-3.8B}}$	0.817	0.580	0.678	0.966	0.665	0.891
Our method$_{\text{Q8B+G12B+P-7B}}$	0.861	0.614	0.717	0.974	0.719	0.914
Our method$_{\text{Q14B+G27B+P-27B}}$	0.879	0.637	0.739	0.982	0.730	0.926
Our method$_{\text{w/o MLs}}$	0.769	0.342	0.477	0.963	0.509	0.845
Our method$_{\text{w/o SF-FD}}$	0.374	0.589	0.459	0.883	0.428	0.810
Our method$_{\text{w/o MLs + SF-FD}}$	0.340	0.290	0.312	0.887	0.315	0.762
Our method	**0.947**	0.702	**0.803**	**0.986**	0.729	**0.985**

Table 4: Main results of concept extraction. Q4B/Q8B/Q14B = Qwen3-4B/Qwen3-8B/Qwen3-14B, G4B/G12B/G27B = Gemma3-4B/Gemma3-12B/Gemma3-27B, P-3.8B/P-7B/P-27B = Phi-3-mini/Phi-3-small/Phi-3-medium, DE = direct extraction, SF-FD = self-feedback, MLs = multiple LLMs.

iterations $iter$ and the integration confidence α used in concept extraction.

Evaluation on Concept Extraction

The main results, ablation study, impact of different scale LLMs and error analysis provide a comprehensive evaluation of concept extraction from multiple insights.

Main Results As shown in Table 4, our proposed method achieves state-of-the-art performance and surpasses all baseline models, attaining the highest F1-score (0.803), ACC (0.986), and AUC (0.985). Notably, the baselines often exhibit a trade-off between precision and recall; for instance, Decomposed-QA achieves high precision (0.935) at the cost of recall (0.570), while GPT-NER shows the inverse. Our method demonstrates a superior balance, leading to the best overall performance.

Ablation Study As shown in Table 4, the ablation study validates the contribution of each component in our framework. Removing the multiple LLMs (w/o MLs) substantially degrades recall, whereas ablating the self-feedback mechanism (w/o SF-FD) causes a sharp drop in precision. The most severe performance degradation is observed when both components are removed, with the F1-score plummeting to 0.312. These findings underscore the synergistic and indispensable roles of both stages in achieving the final robust performance.

Impact of Different Scale LLMs on Concept Extraction As shown in Table 4, our framework's effectiveness is validated by its performance with different scale LLMs. Notably, when paired with the Q14B+G27B+P-27B combination, it surpasses several baseline methods on key metrics (F1 = 0.741, ACC = 0.982), demonstrating its superior ability to extract concepts. Despite this advantage, a significant

Iteration	Confidence	Precision	Recall	F1-score	ACC	MCC	AUC
1	0.5	0.783	0.622	0.693	0.935	0.596	0.939
1	0.7	0.851	0.592	0.698	0.941	0.611	0.947
1	0.9	0.911	0.513	0.657	0.938	0.582	0.952
3	0.5	0.821	0.685	0.746	0.965	0.665	0.967
3	0.7	0.915	0.650	0.760	0.971	0.680	0.972
3	0.9	0.952	0.581	0.721	0.968	0.652	0.965
5	0.5	0.854	**0.741**	0.792	0.981	0.715	0.983
5	0.7	0.947	0.702	**0.803**	**0.986**	**0.729**	**0.985**
5	0.9	**0.975**	0.620	0.758	0.978	0.690	0.983

Table 5: Impact of Iterations and Confidence Thresholds on Performance.

Parameter	**Value**
temperature	0.0
max_length	1000
top_p	0.1
frequency_penalty	0.5
seed	42
iter	5
α	0.7

Table 6: Hyperparameters.

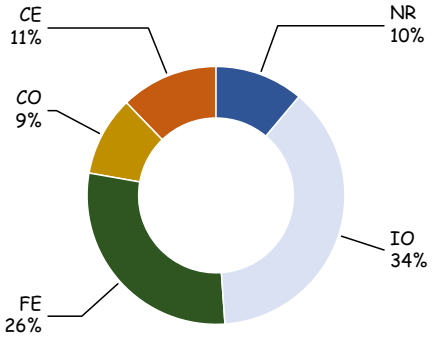

Figure 5: Error distribution. FE = Fomat Error, IO = Invalid Output, CE = Concept Error, CO = Concept Omission, NR = No Response.

performance gap ($\Delta F_1 = 0.062$) persists compared to our primary method using GPT-4o (F1 = 0.803). This starkly illustrates that the superior semantic understanding and reasoning abilities of frontier models are not just beneficial, but indispensable for achieving state-of-the-art performance.

Impact of Different Iterations and Confidence Thresholds on Concept Extraction. We investigate the impact of varying iteration counts ($iter$) and confidence thresholds (α) on concept extraction performance. As shown in Table 5, lower thresholds improve recall, while higher ones emphasize precision, often at the expense of balance. A moderate confidence level yields the best trade-off. Overall, optimal performance is achieved with sufficient iterations ($iter = 5$) and a carefully tuned threshold ($\alpha = 0.7$).

Error Analysis Figure 5 reveals that a low rate of semantic errors, namely Concept Error (CE) and Concept Omission (CO), which demonstrates that the proposed method effectively stimulates the semantic reasoning of LLMs, significantly boosting the accuracy of concept extraction.

	CLIP	**BLIP**	**X-VLM**	**VQA**	**Human**
Score	0.83	0.76	0.81	0.75	0.73

Table 7: Multimodal alignment evaluation. CLIP = CLIP Score, BLIP = BLIP ITM Score, X-VLM = X-VLM ITM Score, VQA = VQA Score.

Evaluation on Multimodal Alignment

Since concept-image alignment uniquely requires semantic and visual reasoning rather than relying on data structure, the evaluation focuses on quantitatively analyzing the semantic consistency between concepts and images. We assess alignment quality using four automatic metrics and human ratings from five experts. Experts score the alignment on a scale from 0 (misaligned) to 1 (perfect). As detailed in Table 7, human evaluations yield a score of 0.73, with strong automatic evaluation scores (e.g., CLIP Score = 0.83), confirming the effectiveness of our alignment.

Conclusion

In this paper, we introduce an automatic multimodal educational knowledge graph framework and construct SciMKG. We also present methods for concept extraction and multimodal alignment. The proposed concept extraction method outperforms all SOTA baselines, and its performance is further validated via a study on the impact of LLM scale and detailed error analysis. The proposed modality alignment is also rigorously evaluated through both automatic and human assessments.

Acknowledgments

This work was supported by the National Natural Science Foundation of China (No. 62276026).

References

Agrawal, G.; Pal, K.; Deng, Y.; Liu, H.; and Chen, Y.-C. 2024. Cyberq: Generating questions and answers for cybersecurity education using knowledge graph-augmented llms. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 23164–23172.

Aytekin, M. C.; Saygın, Y.; et al. 2024. ACE: AI-Assisted Construction of Educational Knowledge Graphs with Prerequisite Relations. *Journal of Educational Data Mining*, 16(2): 85–114.

Billsberry, J.; and Alony, I. 2024. The MOOC post-mortem: Bibliometric and systematic analyses of research on massive open online courses (MOOCs), 2009 to 2022. *Journal of Management Education*, 48(4): 634–670.

Bin, Q.; Zuhairi, M. F.; Morcos, J.; and Zhengqiu, L. 2025. A Ontology Construction Method of Course Knowledge Graph Based on Dependency of Knowledge Points. In *2025 19th International Conference on Ubiquitous Information Management and Communication (IMCOM)*, 1–6.

Chen, H.; Shen, X.; Lv, Q.; Wang, J.; Ni, X.; and Ye, J. 2024. SAC-KG: Exploiting Large Language Models as Skilled Automatic Constructors for Domain Knowledge Graph. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 4345–4360. Association for Computational Linguistics.

Chen, P.; Lu, Y.; Zheng, V. W.; Chen, X.; and Yang, B. 2018. KnowEdu: A System to Construct Knowledge Graph for Education. *IEEE Access*, 6: 31553–31563.

Cui, Y.; Che, W.; Liu, T.; Qin, B.; and Yang, Z. 2021. Pre-Training With Whole Word Masking for Chinese BERT. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29: 3504–3514.

Gong, B.; Tan, S.; Feng, Y.; Xie, X.; Li, Y.; Chen, C.; Zheng, K.; Shen, Y.; and Zhao, D. 2024. UKnow: A Unified Knowledge Protocol with Multimodal Knowledge Graph Datasets for Reasoning and Vision-Language Pre-Training. In *Advances in Neural Information Processing Systems*, volume 37, 9612–9633. Curran Associates, Inc.

Jhajj, G.; Zhang, X.; Gustafson, J. R.; Lin, F.; and Lin, M. P.-C. 2024a. Educational Knowledge Graph Creation and Augmentation via LLMs. In *Generative Intelligence and Intelligent Tutoring Systems*, 292–304. Springer Nature Switzerland.

Jhajj, G.; Zhang, X.; Gustafson, J. R.; Lin, F.; and Lin, M. P.-C. 2024b. Educational Knowledge Graph Creation and Augmentation via LLMs. In *Generative Intelligence and Intelligent Tutoring Systems*, 292–304. Springer Nature Switzerland.

Jin, C.; Wu, Y.; Cao, J.; Xiang, J.; Kuo, Y.-L.; Hu, Z.; Ullman, T.; Torralba, A.; Tenenbaum, J.; and Shu, T. 2024. MMToM-QA: Multimodal Theory of Mind Question Answering. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 16077–16102. Association for Computational Linguistics.

Karakoç Öztürk, B. 2021. Digital reading and the concept of ebook: Metaphorical analysis of preservice teachers' perceptions regarding the concept of ebook. *Sage Open*, 11(2): 21582440211016841.

Lan, J.; Li, J.; Wang, B.; Liu, M.; Wu, D.; Wang, S.; and Qin, B. 2024. NLP-AKG: Few-Shot Construction of NLP Academic Knowledge Graph Based on LLM. *arXiv preprint arXiv:2502.14192*.

Lerner, P.; Ferret, O.; and Guinaudeau, C. 2024. Cross-Modal Retrieval for Knowledge-Based Visual Question Answering. In *Advances in Information Retrieval*, 421–438. Springer Nature Switzerland.

Li, I.; Fabbri, A. R.; Tung, R. R.; and Radev, D. R. 2019. What Should I Learn First: Introducing LectureBank for NLP Education and Prerequisite Chain Learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01): 6674–6681.

Li, J.; Li, D.; Xiong, C.; and Hoi, S. 2022a. BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, 12888–12900. PMLR.

Li, J.; Yang, P.; Yue, K.; Duan, L.; and Huang, Z. 2025. Finding Associative Entities in Knowledge Graph by Incorporating User Behaviors. *Journal of Database Management*, 36(1): 1–24.

Li, N.; Shen, Q.; Song, R.; Chi, Y.; and Xu, H. 2022b. MEduKG: A Deep-Learning-Based Approach for Multi-Modal Educational Knowledge Graph Construction. *Information*, 13(2).

Lin, Z.; Pathak, D.; Li, B.; Li, J.; Xia, X.; Neubig, G.; Zhang, P.; and Ramanan, D. 2025. Evaluating Text-to-Visual Generation with Image-to-Text Generation. In *Computer Vision – ECCV 2024*, 366–384. Springer Nature Switzerland.

Lu, M.; Wang, Y.; Yu, J.; Du, Y.; Hou, L.; and Li, J. 2023. Distantly Supervised Course Concept Extraction in MOOCs with Academic Discipline. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 13044–13059. Association for Computational Linguistics.

Ma, Q.; Zhang, M.; Tang, Y.; and Huang, Z. 2023. Att-Sinkhorn: Multimodal Alignment with Sinkhorn-based Deep Attention Architecture. In *2023 28th International Conference on Automation and Computing (ICAC)*, 1–6.

Ma, S.; Xu, C.; Jiang, X.; Li, M.; Qu, H.; Yang, C.; Mao, J.; and Guo, J. 2024. Think-on-Graph 2.0: Deep and Faithful Large Language Model Reasoning with Knowledge-guided Retrieval Augmented Generation. *arXiv preprint arXiv:2407.10805*.

Madaan, A.; Tandon, N.; Gupta, P.; Hallinan, S.; Gao, L.; Wiegreffe, S.; Alon, U.; Dziri, N.; Prabhumoye, S.; Yang, Y.; Gupta, S.; Majumder, B. P.; Hermann, K.; Welleck, S.;

Yazdanbakhsh, A.; and Clark, P. 2023. Self-Refine: Iterative Refinement with Self-Feedback. In *Advances in Neural Information Processing Systems*, volume 36, 46534–46594. Curran Associates, Inc.

Moás, P. M.; and Lopes, C. T. 2023. Automatic Quality Assessment of Wikipedia Articles—A Systematic Literature Review. *ACM Comput. Surv.*, 56(4).

Neira-Maldonado, P.; Quisi-Peralta, D.; Salgado-Guerrero, J.; Murillo-Valarezo, J.; Cárdenas-Arichábala, T.; Galan-Mena, J.; and Pulla-Sanchez, D. 2024. Intelligent Educational Agent for Education Support Using Long Language Models Through Langchain. In *Information Technology and Systems*, 258–268. Springer Nature Switzerland.

Pan, L.; Wang, X.; Li, C.; Li, J.; and Tang, J. 2017. Course Concept Extraction in MOOCs via Embedding-Based Graph Propagation. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 875–884. Asian Federation of Natural Language Processing.

Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, 8748–8763. PMLR.

Song, S.; Lew, J.; Jang, H.; and Yoon, S. 2024. Unsupervised Homography Estimation on Multimodal Image Pair via Alternating Optimization. In *Advances in Neural Information Processing Systems*, volume 37, 61306–61327. Curran Associates, Inc.

Sun, J.; and Zhang, Z. 2024. LLM4EduKG: LLM for Automatic Construction of Educational Knowledge Graph. In *2024 International Conference on Networking and Network Applications (NaNA)*, 269–275.

Wallace, B.; Dang, M.; Rafailov, R.; Zhou, L.; Lou, A.; Purushwalkam, S.; Ermon, S.; Xiong, C.; Joty, S.; and Naik, N. 2024. Diffusion Model Alignment Using Direct Preference Optimization. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 8228–8238.

Wang, S.; Sun, X.; Li, X.; Ouyang, R.; Wu, F.; Zhang, T.; Li, J.; and Wang, G. 2023a. GPT-NER: Named Entity Recognition via Large Language Models. arXiv:2304.10428.

Wang, X.; Meng, B.; Chen, H.; Meng, Y.; Lv, K.; and Zhu, W. 2023b. TIVA-KG: A Multimodal Knowledge Graph with Text, Image, Video and Audio. In *Proceedings of the 31st ACM International Conference on Multimedia*, MM '23, 2391–2399. Association for Computing Machinery.

Wang, X.; Wei, J.; Schuurmans, D.; Le, Q.; Chi, E.; Narang, S.; Chowdhery, A.; and Zhou, D. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Wang, Y.; Lipka, N.; Rossi, R. A.; Siu, A.; Zhang, R.; and Derr, T. 2024. Knowledge graph prompting for multi-document question answering. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, 19206–19214.

Wu, I.; Jayanthi, S.; Viswanathan, V.; Rosenberg, S.; Pakazad, S. K.; Wu, T.; and Neubig, G. 2024. Synthetic Multimodal Question Generation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, 12960–12993. Association for Computational Linguistics.

Xie, T.; Li, Q.; Zhang, J.; Zhang, Y.; Liu, Z.; and Wang, H. 2023. Empirical Study of Zero-Shot NER with ChatGPT. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 7935–7956. Association for Computational Linguistics.

Xie, T.; Li, Q.; Zhang, Y.; Liu, Z.; and Wang, H. 2024. Self-Improving for Zero-Shot Named Entity Recognition with Large Language Models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, 583–593. Association for Computational Linguistics.

Yu, J.; Lu, M.; Zhong, Q.; Yao, Z.; Tu, S.; Liao, Z.; Li, X.; Li, M.; Hou, L.; Zheng, H.-T.; Li, J.; and Tang, J. 2023. MoocRadar: A Fine-grained and Multi-aspect Knowledge Repository for Improving Cognitive Student Modeling in MOOCs. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '23, 2924–2934. Association for Computing Machinery.

Yu, J.; Luo, G.; Xiao, T.; Zhong, Q.; Wang, Y.; Feng, W.; Luo, J.; Wang, C.; Hou, L.; Li, J.; Liu, Z.; and Tang, J. 2020. MOOCCube: A Large-scale Data Repository for NLP Applications in MOOCs. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 3135–3142. Association for Computational Linguistics.

Yu, J.; Wang, Y.; Zhong, Q.; Luo, G.; Mao, Y.; Sun, K.; Feng, W.; Xu, W.; Cao, S.; Zeng, K.; Yao, Z.; Hou, L.; Lin, Y.; Li, P.; Zhou, J.; Xu, B.; Li, J.; Tang, J.; and Sun, M. 2021. MOOCCubeX: A Large Knowledge-centered Repository for Adaptive Learning in MOOCs. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, CIKM '21, 4643–4652. Association for Computing Machinery.

Zeng, Y.; Zhang, X.; and Li, H. 2022. Multi-Grained Vision Language Pre-Training: Aligning Texts with Visual Concepts. In *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, 25994–26009. PMLR.

Zhang, Z.; Zhao, Y.; Gao, H.; and Hu, M. 2024. LinkNER: Linking Local Named Entity Recognition Models to Large Language Models using Uncertainty. In *Proceedings of the ACM Web Conference 2024*, WWW '24, 4047–4058. Association for Computing Machinery.

Zhou, W.; Zhang, S.; Gu, Y.; Chen, M.; and Poon, H. 2023. Universalner: Targeted distillation from large language models for open named entity recognition. *arXiv preprint arXiv:2308.03279*.