MOOC-MESE: Dataset and Framework for Multimodal Entity Set Expansion in MOOCs

Shichao Yue[1] **Jing Wan**[1,*] **Yubin Chen**[1]
[1]Beijing University of Chemical Technology
2025210643@mail.buct.edu.cn, wanj@mail.buct.edu.cn, 455375842@qq.com

Abstract

Entity Set Expansion (ESE) aims to expand entities belonging to the same semantic category given only a small set of seed entities. Traditional ESE methods are primarily text-based and struggle to address two key challenges in MOOC scenarios: (1) Multimodal Semantic Dependency; (2) Generalization Issues. To facilitate research in this area, we construct MOOC-MESE, the first Chinese multimodal ESE dataset for MOOCs, which contains over 45k aligned image-text pairs. The dataset provides rich multimodal semantic information for entity expansion research in educational scenarios. Furthermore, we propose GMRM (Generative Multimodal Retrieval Model), a multimodal ESE framework that reformulates the task as a multimodal generative retrieval problem. By achieving multimodal alignment and fully exploiting cross-modal semantics, GMRM effectively mitigates generalization issues. Extensive experiments and analyses on MOOC-MESE demonstrate both the high quality of the dataset and the effectiveness of GMRM. The code is publicly available at: https://github.com/constantinus-xi/mooc-mese.

1 Introduction

Entity set expansion (ESE) aims to discover new entities belonging to the same semantic category as a small set of seed entities from a candidate vocabulary (Huang et al., 2024; Li et al., 2023d). For example, given the seed entities "Queue", "AVL Tree", and "Stack", the goal is to extend the category with other data structures such as "B+ Tree", "Red-Black Tree", and "Graph". Entity set expansion (ESE) plays an important role in knowledge mining by automatically discovering new entities that belong to the same semantic category. Expanded entity sets have been shown to benefit a wide range of downstream NLP and IR applications, including web search (Chen et al., 2016), question answering (Wang and Cohen, 2008), query understanding (Hua et al., 2017), taxonomy construction (Li et al., 2023b; Velardi et al., 2013; Shen et al., 2018a), knowledge graph construction (Li et al., 2023a,c; Xu et al., 2023), and semantic search (Xiong et al., 2017; Shen et al., 2018b).

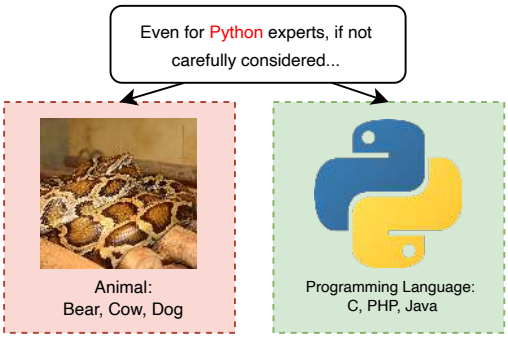


Figure 1: An example of semantic ambiguity in entity recognition caused by insufficient contextual information. The term “Python” may refer to either a programming language or an animal, leading to different semantic interpretations even for human experts when multimodal or contextual cues are absent.

Traditional entity set expansion methods are predominantly text-based and typically rely on sparse contextual information and limited semantic representations, making them insufficient for modeling entity semantics in complex scenarios. This limitation becomes more pronounced in MOOCs environments. Specifically, text-only ESE methods face the following challenges in MOOCs:

1) **Multimodal Semantic Dependency.** Course semantics are jointly conveyed through textual and visual content. Unlike social media or textbook scenarios, instructional content in MOOCs heavily depends on the interaction between text and images and is often accompanied by colloquial expressions,

*Corresponding author

Findings of the Association for Computational Linguistics: IJCNLP-AACL 2026, pages 4249–4258
November 6-10, 2026 ©2026 Association for Computational Linguistics

resulting in inherent semantic incompleteness and noise in text-only representations,as shown in Figure 1.

2) **Generalization Issues.** MOOC platforms cover a large number of courses from diverse disciplines and professional fields, involving extensive and highly heterogeneous domain-specific terminology. Therefore, compared with single-domain or social media scenarios, entity categories are more complex and exhibit broader semantic diversity. Such cross-domain and diverse entity distributions impose higher requirements on the model's generalization ability.

To address the limitations of unimodal entity set expansion in the MOOC domain, we extend MOOC-ESE with 45,938 images and construct MOOC-MESE, the first Chinese multimodal ESE dataset for MOOCs, thereby alleviating the lack of multimodal information in existing resources.

Furthermore, we propose GMRM (Generative Multimodal Retrieval Model), a generative retrieval framework for multimodal entity set expansion. It reformulates ESE as a multimodal generative retrieval problem by assigning discrete identifiers to entity-related images and modeling the associations among context, entities, images, and identifiers. GMRM consists of a training stage and an expansion stage. The training stage follows a two-step strategy: learning the mapping between images and identifiers, and retrieving image identifiers from textual descriptions. This design enables end-to-end multimodal alignment, supports flexible expansion through incremental training, decouples entity representation from entity expansion, and improves generalization through prompting and an iterative expansion process.

In summary, the main contributions are as follows:

- We define the task of Multimodal Entity Set Expansion (MESE) in the MOOC scenario and construct a new MESE dataset named MOOC-MESE.
- To address semantic ambiguity and cross-domain generalization issues in multimodal entity expansion, we propose a strong multimodal framework, GMRM, which reformulates the multimodal entity set expansion task as a generative retrieval problem.
- Extensive experiments on MOOC-MESE demonstrate that our method achieves the best MAP performance, verifying the effectiveness of GMRM.

2 Related Work

2.1 Entity Set Expansion.

Existing ESE methods are typically built upon powerful text-based pre-trained language models, such as BERT (Devlin et al., 2019), BERT-Chinese, and RoBERTa-wwm (Cui et al., 2021), which commonly serve as backbone encoders for entity representation or prompt-based generation. On this basis, research on entity set expansion mainly follows two paradigms: generative methods and retrieval-based methods (Kardos et al., 2025).

Generative methods exploit the parametric knowledge stored in pre-trained language models. One line of work directly generates candidate entities through carefully designed prompts or probing strategies (Zhang et al., 2020). Another line of work adopts category-guided generation, explicitly modeling category semantics during generation to further improve expansion quality, such as CGExpan (Zhang et al., 2020). These approaches are lightweight and flexible, and perform well in purely textual scenarios. However, their effectiveness heavily depends on the internal knowledge of language models, making them less suitable for scenarios involving multimodal entities or fine-grained visual distinctions.

Retrieval-based methods encode entities from their textual contexts and rank candidates by their similarity to seed entities. RetExpan (Li et al., 2025) first retrieves candidates using positive seeds and then re-ranks them using negative seeds. In multimodal settings, existing methods commonly adopt dual-encoder retrieval architectures that encode textual and visual inputs separately and fuse them through cross-modal interactions, such as CLIP (Radford et al., 2021) and ALBEF (Li et al., 2021). However, these methods often rely on large-scale pre-training or fine-tuning, making them costly to adapt to new expansion targets or domains.

To address these limitations, we build upon GRACE (Li et al., 2024b), which uses OpenFlamingo (Awadalla et al., 2023) as its backbone, and propose a generative retrieval-based framework with an iterative expansion process. Different from traditional dense retrieval methods (Li et al., 2023e; Gillick et al., 2018; Karpukhin et al., 2020; Ni et al., 2022), it provides an end-to-end solution (Cao et al.,

2020) that better exploits the memorization and reasoning abilities of large language models (Bevilacqua et al., 2022; Sun et al., 2023), thereby improving cross-modal alignment and enabling scalable expansion through incremental training.

2.2 Multi-Modal Datasets.

With the rapid development of vision-language models, large-scale multimodal datasets have become increasingly important for supporting cross-modal representation learning and downstream multimodal tasks. Existing datasets are commonly constructed from web-scale image-text resources or through automatic annotation pipelines. For example, Thomee et al. (2016) collected metadata of approximately 99 million Flickr images, including tags and geographic information, to build YFCC100M. Similarly, the SBU Captioned Photo Dataset (Ordonez et al., 2011) and Conceptual Captions (Sharma et al., 2018) extracted image descriptions from raw HTML alt-text and applied predefined filtering rules to obtain cleaner image-text pairs.

Recent studies further improve dataset quality and domain specificity by introducing model-assisted annotation, expert models, or human verification. Ruan et al. (2025) combined large language models with human verification to construct a dataset containing images and instruction-following dialogues. MedTrinity-25M (Xie et al., 2025) employed an automated pipeline with domain-specific expert models to build a large-scale medical multimodal dataset containing over 25 million images. In the field of entity set expansion, MESED (Li et al., 2024a) constructed a large-scale, fine-grained, manually calibrated multimodal ESE dataset by scoring the relevance between images and entities. These studies show that multimodal dataset construction has gradually shifted from simple web-scale collection to more task-oriented and quality-controlled construction pipelines.

However, existing datasets are mostly built on general-domain corpora such as Wikipedia and largely rely on textual information, which limits their applicability to educational scenarios. In MOOCs, entity semantics are often conveyed through both text and visual content, while existing MOOC datasets lack multimodal annotations. To fill this gap, we construct MOOC-MESE based on MOOCCubeX (Yu et al., 2021) and MOOC-ESE (Chen et al., 2024), providing the first Chinese multimodal ESE benchmark for the MOOC domain.

3 Task Definition

Given a seed entity set $S = \{e_1, e_2, \ldots, e_k\}$ and a candidate vocabulary V, multimodal entity set expansion (MESE) aims to retrieve entities from V that belong to the same semantic category as the seeds. Each entity e_i is paired with a textual context t_i and an aligned image v_i, and the corpus is defined as

$$D = \{(e_i, t_i, v_i)\}_{i=1}^{|D|}. \quad (1)$$

The goal is to rank candidate entities by their multimodal semantic consistency with S.

4 Dataset Construction

Due to the lack of publicly available multimodal entity set expansion datasets, we built a Chinese-language dataset for the MOOC domain, named MOOC-MESE (Multi-modal Entity Set Expansion in MOOCs), based on the MOOCCubeX (Yu et al., 2021) dataset, which differs from existing datasets by explicitly incorporating aligned multimodal information and providing annotated entity sets tailored for the ESE task. The construction process of MOOC-MESE mainly consists of five steps: data collection, data preprocessing, entity set expansion, keyword recognition, and multimodal information supplementation. In the multimodal supplementation stage, this study aims to supplement each entity-sentence pair with corresponding visual modality information. Considering that the same entity may have different meanings in different contexts, we design and implement a new processing workflow. This workflow can effectively extract key terms from different contexts and introduce an image filtering mechanism to ensure the relevance and quality of the supplemented images.

The construction process of the MOOC-MESE dataset includes the following key stages:

Data Collection. This study uses the MOOCCubeX dataset as the base corpus, and through a systematic data preprocessing pipeline, two types of normalized data collections are constructed: course-entity association data and entity-context correspondence data. According to statistics, the final constructed corpus contains 476 online course resources, and a total of 45,938 entity instances with complete contextual information were extracted and annotated.

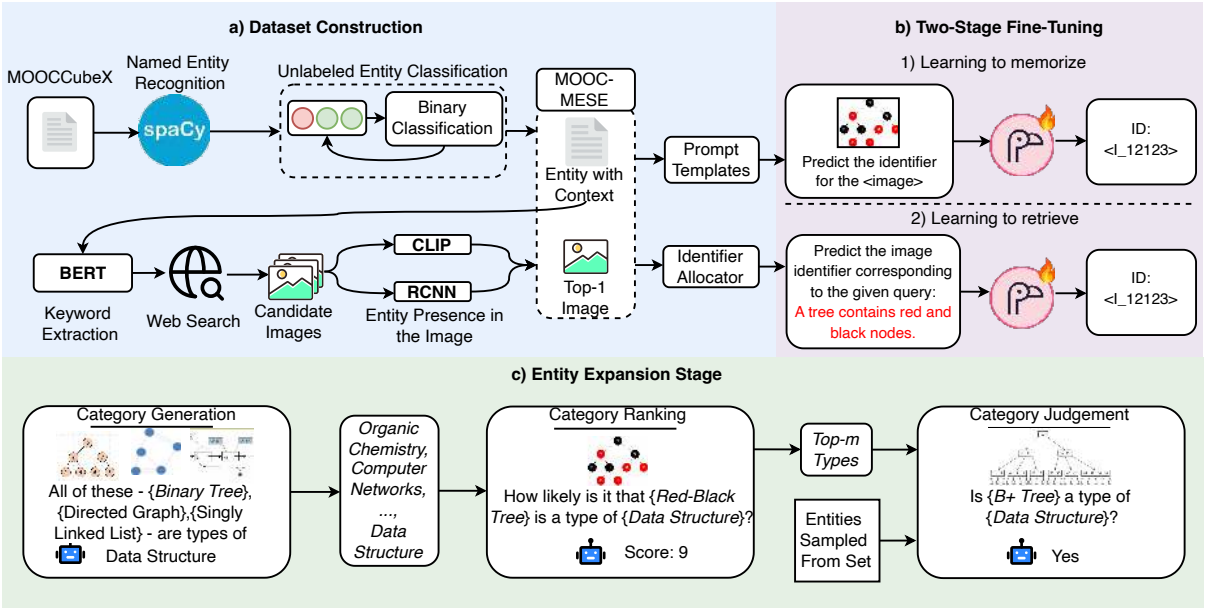


Figure 2: Overview of the proposed multimodal entity set expansion framework.

Data Preprocessing. We use the spaCy natural language processing tool to perform entity recognition and annotation on the raw text in the MOOCCubeX dataset. This tool automatically detects entity objects in the text through its Named Entity Recognition (NER) module and assigns a globally unique identifier to each recognized entity, while precisely recording the position information of the entity in the context. After this process, a structured corpus is constructed, in which each entity is annotated with complete contextual position information, providing high-quality input for the subsequent pretraining of the masked language model.

Unlabeled Entity Classification. Due to the size limitation of the annotated data provided by the MOOCCubeX dataset, traditional evaluation methods based on fixed standard entity sets are inadequate for fully reflecting model performance. Therefore, we design a validation mechanism based on a classification model: First, an initial standard entity set is constructed, then a binary classification model is used to assess whether the candidate entities output by the system belong to the target entity set, and finally, the validated entities are dynamically added to the standard set. This iterative evaluation strategy not only addresses the issue of insufficient annotated data, but also continuously optimizes the completeness of the evaluation benchmarks.

Keyword Retrieval. BERT is used to extract key terms from the entity context, then concatenate the keywords with the corresponding entity, and use web search to generate search images. Ten images are retrieved for each keyword.

Entity Image Re-ranking. From the 10 images retrieved, one image needs to be selected as the image modality for the entity. An ideal image should reflect the content of the sentence while including the entity of the sentence. Considering both aspects, a simple and effective image re-ranking algorithm is designed to select the most suitable image v for sentence t and entity e. The score for the image-text pair is shown in Equation (2):

$$\begin{aligned} \text{score}(v_i, t, e) = \lambda \cdot \widetilde{\text{CLIP}}(v_i, t) \\ + (1 - \lambda) \cdot \widetilde{\text{DEC}}(v_i, e) \end{aligned} \tag{2}$$

where the CLIP model is used to measure the relevance between the image and text, and the DEC model employs Faster-CNN to detect whether corresponding entities are present in the image. Table 1 summarizes the detailed statistical information of the MOOC-MESE dataset. These statistics indicate that MOOC-MESE provides sufficient data diversity and scale to support robust training and evaluation for multimodal entity set expansion tasks.

5 Methodology

5.1 Overview

We describe the proposed GMRM framework for MESE, inspired by GRACE (Li et al., 2024b),

Table 1: Comparison of existing ESE datasets.

Metric	Wiki	APR	CoNLL	ONs	MESED	MOOC-MESE
Queries	5	5	1	1	10	1
Seed	3	3	10	10	3/5	3
Entities	33K	76K	6K	20K	14K	45K
Sentences	973K	1043K	21K	144K	434K	64K
Multi-Modal	N	N	N	N	Y	Y

which reformulates multimodal entity set expansion as a generative expansion task. We divide GMRM into two stages: the multimodal large language model training stage and the entity set expansion stage. In the first stage, we design an identifier-based generative cross-modal retrieval strategy to train the memorization and retrieval abilities of the multimodal large language model. In the second stage, we design a generative paradigm based on prompt engineering to dynamically expand the entity set, thereby improving the generalization of the model. We note that GMRM is proposed to provide a robust multimodal baseline and to explore the effectiveness of different pretraining tasks.

5.2 Multimodal Language Model Training Phase

In the model training phase, we adopt a generative retrieval approach. Specifically, we first design image identifiers and then train the model's memory and retrieval capabilities. Implementation details are provided in the following sections.

5.2.1 Multiple Identifiers

To enable generative retrieval of visual inputs, each image needs to be mapped to a discrete and recognizable model identifier. These identifiers serve as a bridge between the visual modality and the language model, allowing images to be retrieved within a unified token space. In this work, we explore three types of image identifiers: string identifiers, digit identifiers, and atomic identifiers. Among them, string and digit identifiers may be split into multiple sub-tokens by the tokenizer, which can weaken the consistency and uniqueness of image representations. Therefore, we adopt atomic identifiers, where each image is assigned a dedicated special token to ensure a unique and indivisible semantic representation.

5.2.2 Model Memory Capability

In this stage, the model learns the mapping from images to their corresponding identifiers, thus facilitating the formation of visual memory. Furthermore, the model can generate identifiers as retrieval results instead of generating actual images. Each image in the dataset is assigned a unique identifier and is represented as a sequence of tokens. Then, a multimodal language model is trained to encode the image information in its parameters. For each image i in the dataset, the trained model associates it with its corresponding identifier. This process can be formally represented by the following Equation (3):

$$L_m = \text{MLLM}(i; \text{inst}_m) \tag{3}$$

where inst_m is a text instruction obtained through the prompt generator.

5.2.3 Model Retrieval Capability

For masked language models, it is insufficient to store image information only within the model parameters. The model must possess the ability to accurately retrieve and recall the corresponding image in response to user queries. This capability aligns with the design concept of entity set expansion, making it crucial for effective multimodal entity set expansion. To achieve this goal, this section proposes a query-driven image identifier prediction method.

Given the user query inst_m, the model needs to learn a mapping function $f : q \rightarrow i$, where $i \in I$ represents the corresponding image identifier (with i being the identifier space). The objective of optimizing this mapping function is to minimize the difference between the predicted identifier and the true identifier, ensuring that the model can accurately associate the query with the target image. The mathematical formalization of this process is described by the following Equation (4):

$$L_r = \text{MLLM}(q; \text{inst}_r) \tag{4}$$

where inst_r is a text instruction for predicting the image identifier corresponding to the given query.

5.3 Entity Expansion Phase

The research method in this phase draws inspiration from the generative iterative frameworks in the existing literature, with key innovations introduced. Compared to previous studies, this section innovatively introduces a category-guided mechanism based on the image modality of the seed entity set. Specifically, this mechanism generates a multimodal integrated category representation by combining the text and image modalities of the seed entities. Based on the integrated multimodal

category, an iterative framework is designed to optimize the entity expansion process.

The core idea of the framework is to iteratively optimize the integrated multimodal category and expand the set of candidate entities. Specifically, in each iteration, new candidate entities are first generated based on the current multimodal integrated category, and then these candidate entities are used to further optimize the category representation. The newly optimized category is subsequently used for further entity set expansion.

5.3.1 Multimodal Language Model-Based Integrated Category Generation

The category name generation module is designed to generate a set of high-quality category names based on the input seed entity set and its image modality. The core mechanism of this module is to automatically construct category probing queries and iteratively query the pre-trained multimodal language model from the previous stage to obtain candidate category names. However, this direct query approach may lead to the generation of overly coarse-grained category names that fail to fully reflect the semantic features of the seed entities. To address this issue, we propose a category optimization strategy based on an iterative framework: in each iteration, a new prompt template is reconstructed by concatenating the currently generated category.

This dynamic prompt construction method effectively guides the model to generate more precise and polysemous categories. Through multiple iterations, the results of the category name generation can be progressively refined, ultimately producing a polysemous category that accurately reflects the semantic features of the seed entities.

5.3.2 Iterative Expansion Framework Based on Integrated Categories

In this section, we use the multimodal integrated category generation module to iteratively obtain the semantic categories of the iteration set and expand the entity set based on these semantic categories. To measure the similarity between entities and semantic categories, we adopt the metric proposed by Zhang et al. (2020), as shown in Equation (5):

$$M^k(c,e) = \frac{1}{k} \max_{S \subset X_e, |S|=k} \sum_{x \in S} \max_{x' \in X_c} \cos(x, x') \tag{5}$$

Given a category name c, six entity queries are constructed in the six Hearst patterns, and a pre-trained multimodal language model is queried to obtain the set of six token embeddings, denoted as X_c. Additionally, X_e is used to represent the set of all contextualized representations of entity e in the given corpus.

6 Experiments

6.1 Experiment Setup

We evaluate our method on MOOC-MESE and compare it with representative baselines from both text-based and multimodal paradigms. For text-based entity set expansion, we include **BERT** (Devlin et al., 2019), **BERT-Chinese**, and **RoBERTa-wwm** (Cui et al., 2021). To evaluate the contribution of visual information, we compare our method with multimodal pre-trained models, including **CLIP** (Radford et al., 2021) and **ALBEF** (Li et al., 2021).

6.2 Main Result

Table 2 reports the performance comparison of different models in the MOOC-MESE dataset. In general, the proposed multimodal framework consistently outperforms all baseline methods in MAP@10, MAP@20, and MAP@50, demonstrating its effectiveness in multimodal entity set expansion. The proposed multimodal methods based on ViT-L-14 + ChatGLM-6B and ViT-L-14 + InternLM-XComposer-7B achieve the best results. In particular, both models significantly outperform the strongest baseline (ROBERTA-wwm on category), highlighting the advantage of integrating visual information with large-scale Chinese language models. The experimental results show that the multimodal architecture based on generative retrieval can effectively preserve linguistic knowledge, leading to more accurate and robust entity expansion. Overall, these findings validate the effectiveness and practicality of the proposed framework for Chinese MOOC data processing tasks.

6.3 Ablation Studies

To investigate the individual contributions of the memory module and the retrieval module in the proposed framework, we conducted a series of ablation experiments under controlled settings. Specifically, we evaluate four configurations: (1) **Not Fine-tuned**, where neither the memory nor retrieval module is applied and the backbone lan-

Table 2: Comparison of models on MOOC-MESE dataset

Modality	Model	MAP@10	MAP@20	MAP@50
Text	BERT	1.42	0.71	0.29
	BERT-Chinese	16.67	8.33	3.52
	RoBERTa-wwm	67.21	38.09	42.84
	BERT on category	3.21	2.57	0.99
	BERT-Chinese on category	29.44	16.04	6.42
	RoBERTa-wwm on category	87.88	62.36	51.67
Text+Image	CLIP	57.67	55.88	51.02
	ALBEF	66.59	67.72	61.68
	ViT-L-14 + THUDM/chatglm-6b(Ours)	97.54	**96.54**	**95.82**
	ViT-L-14 + internlm-xcomposer-7b(Ours)	**98.23**	93.54	93.32

Table 3: Ablation study results on MOOC-MESE.

Setting	Language Model	MAP@10	MAP@20	MAP@50
Not Fine-tuned	ChatGLM-6B	76.43	67.32	59.22
	XComposer-7B	80.43	71.43	65.43
Memory Only	ChatGLM-6B	91.74	90.18	88.32
	XComposer-7B	**93.71**	89.61	86.12
Retrieval Only	ChatGLM-6B	80.71	72.94	66.02
	XComposer-7B	**84.80**	**76.43**	**77.01**
Fine-tuned	ChatGLM-6B	97.54	**96.54**	**95.82**
	XComposer-7B	**98.23**	93.54	93.32

guage model is used directly; (2) **Memory Module Only**, where only the memory-based component is enabled while the retrieval component is removed; (3) **Retrieval Module Only**, where only the retrieval-based component is applied; and (4) **Fine-tuned**, which corresponds to the full two-stage framework with both memory and retrieval modules enabled. All configurations are evaluated using the same experimental setup and metrics to ensure fair comparison.

Table 3 reports the ablation results on the MOOC-MESE dataset. Compared with the non-fine-tuned setting, enabling either module leads to consistent performance improvements across all MAP metrics, demonstrating the effectiveness of both components. Notably, the Memory Module Only setting yields substantially larger gains than the Retrieval Module Only setting for both backbone models, indicating that the memory module plays a more critical role in capturing and consolidating entity-level semantic information. The Retrieval Module Only configuration also improves performance over the non-fine-tuned baseline, but the improvement margin is relatively limited.

When both modules are jointly applied in the Fine-tuned setting, the model achieves the best performance, significantly outperforming all ablated variants. This result confirms that the memory and retrieval modules are complementary, and their combination enables more effective multimodal

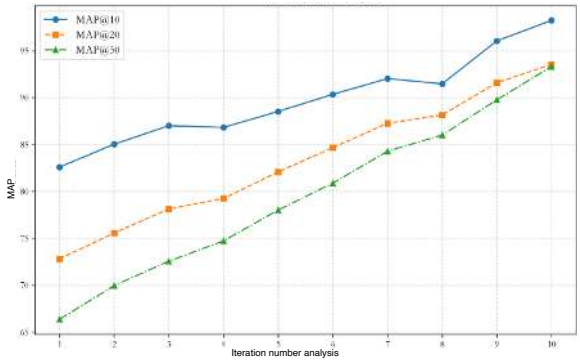


Figure 3: Impact of category-guided iteration numbers on multimodal entity set expansion performance. MAP@10, MAP@20, and MAP@50 show an overall upward trend despite minor fluctuations, reaching the best results at the 10th iteration.

entity set expansion. Overall, the ablation results validate the design of the two-stage framework and highlight the importance of integrating both memory consolidation and retrieval-based reasoning.

6.4 Iteration number analysis

In the iterative expansion framework based on integrated categories, we investigate the effect of the number of category-guided iterations on entity set expansion. As shown in the line chart, MAP@10, MAP@20, and MAP@50 generally improve as the number of iterations increases. In the first five iterations, the performance gains are particularly clear, with MAP@10 increasing from 82.60 to 88.53, MAP@20 from 72.81 to 82.08, and MAP@50 from 66.37 to 78.02. This indicates that early iterations effectively improve entity expansion performance.

Although slight fluctuations occur in some intermediate iterations, such as the temporary decrease of MAP@10 at the fourth and eighth iterations, the overall trend remains upward. These fluctuations may be caused by the introduction of redun-

dant entities or entities with ambiguous semantic boundaries during iterative expansion. At the tenth iteration, all three metrics reach their best results, with MAP@10, MAP@20, and MAP@50 reaching 98.23, 93.54, and 93.32, respectively.

Therefore, considering both performance and computational cost, we set the number of iterations to 10. This experiment demonstrates the effectiveness of the iterative module and shows that progressively introducing integrated category information can improve multimodal entity set expansion.

6.5 Case Study

This section presents a case study to analyze the entity expansion performance in the Data Structure category. As shown in Table 4, given the seed entities "Singly Linked List", "Binary Tree", and "Directed Graph", different methods show clear differences in their expansion results.

The Only Text method expands several entities that are only similar at the surface lexical or terminological level. For example, it tends to include concepts that are related to words such as "tree" or "graph" but do not strictly belong to the Data Structure category. This indicates that the text-only method mainly relies on shallow textual semantics, which can easily lead to semantic drift.

The Multi-Modal method introduces visual information to some extent, but it still expands entities that are visually similar while being closer to application scenarios or related tools of data structures rather than data structures themselves. This suggests that simply incorporating multimodal information is not sufficient to distinguish category concepts from their application forms.

In contrast, GMRM can better integrate textual semantics and visual information. It improves category discrimination and reduces cross-domain expansion. The expanded entities are more semantically consistent with the Data Structure category, demonstrating that the proposed method can generate higher-quality entity expansion results in multimodal scenarios.

Table 4: Case study in the Data Structure category with the seed entities "Singly Linked List", "Binary Tree", and "Directed Graph". Red entities indicate incorrect expansions.

Only Text	Multi-Modal	GMRM
链表	双向链表	循环链表
Linked List	Doubly Linked List	Circular Linked List
双向链表	链表	哈夫曼树
Doubly Linked List	Linked List	Huffman Tree
依赖图	进程树	二叉查找树
Dependency Graph	Process Tree	Binary Search Tree
控制流图	知识图谱	平衡树
Control Flow Graph	Knowledge Graph	Balanced Tree
白川英树	时序图	有向无环图
Shirakawa Hideki	Sequence Diagram	Directed Acyclic Graph
目录树	二叉查找树	红黑树
Directory Tree	Binary Search Tree	Red-Black Tree
循环链表	UML类图	链表
Circular Linked List	UML Class Diagram	Linked List
甘特图	无向图	平衡二叉搜索树
Gantt Chart	Undirected Graph	Balanced Binary Search Tree

7 Conclusion

In this paper, we construct MOOC-MESE, a new entity set expansion dataset that alleviates the lack of multimodal data in the education domain. MOOC-MESE covers entities from multiple professional fields, posing new challenges to the generalization ability of ESE models. Based on this dataset, we propose GMRM, a generative multimodal retrieval framework that reformulates entity set expansion as a generative retrieval problem. Through a two-stage training strategy and a category-guided iterative expansion framework, GMRM achieves multimodal alignment and improves generalization ability. Compared with existing unimodal and multimodal models, GMRM achieves the best performance.

Limitations

Although this study achieved promising experimental results, several limitations remain. First, although the constructed multimodal dataset incorporates image information, the quality of image-text pairs still requires further improvement and more extensive statistical or quantitative evaluation. Second, due to computational resource constraints, the experiments were mainly conducted on a relatively limited-scale Chinese MOOC dataset, and the generalization ability of the framework on broader domains and larger-scale multimodal corpora has not yet been fully validated; the framework has also not yet been compared with recent large language model-based entity set expansion methods. Finally, the current method mainly focuses on static multimodal information and does not model the temporal and interactive contextual dynamics that may exist in real educational scenarios.

References

Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, et al.

2023. Openflamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*.

Michele Bevilacqua, Giuseppe Ottaviano, Patrick Lewis, Scott Yih, Sebastian Riedel, and Fabio Petroni. 2022. Autoregressive search engines: Generating substrings as document identifiers. In *Advances in Neural Information Processing Systems*, volume 35, pages 31668–31683. Curran Associates, Inc.

Nicola De Cao, Gautier Izacard, Sebastian Riedel, and Fabio Petroni. 2020. Autoregressive entity retrieval. *CoRR*, abs/2010.00904.

Yubin Chen, Jing Wan, Yexing Du, Jifan Yu, Lei Hou, and Juanzi Li. 2024. Entity set expansion based on category prompts in moocs. In *Knowledge Science, Engineering and Management*, pages 318–332, Singapore. Springer Nature Singapore.

Zhe Chen, Michael Cafarella, and H. V. Jagadish. 2016. Long-tail vocabulary dictionary extraction from the web. In *Proceedings of the Ninth ACM International Conference on Web Search and Data Mining*, WSDM '16, page 625–634, New York, NY, USA. Association for Computing Machinery.

Yiming Cui, Wanxiang Che, Ting Liu, Bing Qin, and Ziqing Yang. 2021. Pre-training with whole word masking for chinese bert. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 29:3504–3514.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Daniel Gillick, Alessandro Presta, and Gaurav Singh Tomar. 2018. End-to-end retrieval in continuous space. *CoRR*, abs/1811.08008.

Wen Hua, Zhongyuan Wang, Haixun Wang, Kai Zheng, and Xiaofang Zhou. 2017. Understand short texts by harvesting and analyzing semantic knowledge. *IEEE Transactions on Knowledge and Data Engineering*, 29(3):499–512.

Shulin Huang, Shirong Ma, Yangning Li, Yinghui Li, and Hai-Tao Zheng. 2024. From retrieval to generation: Efficient and effective entity set expansion. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management*, CIKM '24, page 921–931, New York, NY, USA. Association for Computing Machinery.

Péter Kardos, András London, and Richárd Farkas. 2025. Human-in-the-loop entity set expansion using knowledge graphs. In *Proceedings of the Workshop on Explainable AI for Knowledge Graphs (XAIKG)*.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation. In *Advances in Neural Information Processing Systems*, volume 34, pages 9694–9705. Curran Associates, Inc.

Yangning Li, Jiaoyan Chen, Yinghui Li, Yuejia Xiang, Xi Chen, and Hai-Tao Zheng. 2023a. Vision, deduction and alignment: An empirical study on multi-modal knowledge graph alignment. In *ICASSP 2023 - 2023 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 1–5.

Yangning Li, Jiaoyan Chen, Yinghui Li, Tianyu Yu, Xi Chen, and Hai-Tao Zheng. 2023b. Embracing ambiguity: Improving similarity-oriented tasks with contextual synonym knowledge. *Neurocomputing*, 555:126583.

Yangning Li, Yinghui Li, Xi Chen, Hai-Tao Zheng, and Ying Shen. 2023c. Active relation discovery: Towards general and label-aware open relation extraction. *Knowledge-Based Systems*, 282:111094.

Yangning Li, Tingwei Lu, Hai-Tao Zheng, Yinghui Li, Shulin Huang, Tianyu Yu, Jun Yuan, and Rui Zhang. 2024a. Mesed: A multi-modal entity set expansion dataset with fine-grained semantic classes and hard negative entities. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pages 8697–8706.

Yangning Li, Qingsong Lv, Tianyu Yu, Yinghui Li, Xuming Hu, Wenhao Jiang, Hai-Tao Zheng, and Hui Wang. 2025. Ultrawiki: Ultra-fine-grained entity set expansion with negative seed entities. In *2025 IEEE 41st International Conference on Data Engineering (ICDE)*, pages 1855–1868.

Yinghui Li, Shulin Huang, Xinwei Zhang, Qingyu Zhou, Yangning Li, Ruiyang Liu, Yunbo Cao, Hai-Tao Zheng, and Ying Shen. 2023d. Automatic context pattern generation for entity set expansion. *IEEE Transactions on Knowledge and Data Engineering*, 35(12):12458–12469.

Yongqi Li, Wenjie Wang, Leigang Qu, Liqiang Nie, Wenjie Li, and Tat-Seng Chua. 2024b. Generative cross-modal retrieval: Memorizing images in multimodal language models for retrieval and beyond. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11851–11861, Bangkok, Thailand. Association for Computational Linguistics.

Yongqi Li, Nan Yang, Liang Wang, Furu Wei, and Wenjie Li. 2023e. Generative retrieval for conversational question answering. *Information Processing Management*, 60(5):103475.

Jianmo Ni, Chen Qu, Jing Lu, Zhuyun Dai, Gustavo Hernandez Abrego, Ji Ma, Vincent Zhao, Yi Luan, Keith Hall, Ming-Wei Chang, and Yinfei Yang. 2022. Large dual encoders are generalizable retrievers. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9844–9855, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Vicente Ordonez, Girish Kulkarni, and Tamara Berg. 2011. Im2text: Describing images using 1 million captioned photographs. In *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc.

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. 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*, pages 8748–8763. PMLR.

Jiacheng Ruan, Wenzhen Yuan, Zehao Lin, Ning Liao, Zhiyu Li, Feiyu Xiong, Ting Liu, and Yuzhuo Fu. 2025. Mm-camobj: A comprehensive multimodal dataset for camouflaged object scenarios. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(7):6740–6748.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, Melbourne, Australia. Association for Computational Linguistics.

Jiaming Shen, Zeqiu Wu, Dongming Lei, Chao Zhang, Xiang Ren, Michelle T. Vanni, Brian M. Sadler, and Jiawei Han. 2018a. Hiexpan: Task-guided taxonomy construction by hierarchical tree expansion. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '18, page 2180–2189, New York, NY, USA. Association for Computing Machinery.

Jiaming Shen, Jinfeng Xiao, Xinwei He, Jingbo Shang, Saurabh Sinha, and Jiawei Han. 2018b. Entity set search of scientific literature: An unsupervised ranking approach. In *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval*, SIGIR '18, page 565–574, New York, NY, USA. Association for Computing Machinery.

Weiwei Sun, Lingyong Yan, Zheng Chen, Shuaiqiang Wang, Haichao Zhu, Pengjie Ren, Zhumin Chen, Dawei Yin, Maarten Rijke, and Zhaochun Ren. 2023. Learning to tokenize for generative retrieval. In *Advances in Neural Information Processing Systems*, volume 36, pages 46345–46361. Curran Associates, Inc.

Bart Thomee, David A. Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. 2016. Yfcc100m: the new data in multimedia research. *Commun. ACM*, 59(2):64–73.

Paola Velardi, Stefano Faralli, and Roberto Navigli. 2013. OntoLearn reloaded: A graph-based algorithm for taxonomy induction. *Computational Linguistics*, 39(3):665–707.

Richard C. Wang and William W. Cohen. 2008. Iterative set expansion of named entities using the web. In *2008 Eighth IEEE International Conference on Data Mining*, pages 1091–1096.

Yunfei Xie, Ce Zhou, Lang Gao, Juncheng Wu, Xianhang Li, Hong-Yu Zhou, Sheng Liu, Lei Xing, James Y Zou, Cihang Xie, and Yuyin Zhou. 2025. Medtrinity-25m: A large-scale multimodal dataset with multigranular annotations for medicine. In *International Conference on Learning Representations*, volume 2025, pages 6036–6060.

Chenyan Xiong, Russell Power, and Jamie Callan. 2017. Explicit semantic ranking for academic search via knowledge graph embedding. In *Proceedings of the 26th International Conference on World Wide Web*, WWW '17, page 1271–1279, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Jin Xu, Yangning Li, Xiangjin Xie, Yinghui Li, Niu Hu, Haitao Zheng, and Yong Jiang. 2023. Investigating graph structure information for entity alignment with dangling cases. *arXiv preprint arXiv:2304.04718*.

Jifan Yu, Yuquan Wang, Qingyang Zhong, Gan Luo, Yiming Mao, Kai Sun, Wenzheng Feng, Wei Xu, Shulin Cao, Kaisheng Zeng, Zijun Yao, Lei Hou, Yankai Lin, Peng Li, Jie Zhou, Bin Xu, Juanzi Li, Jie Tang, and Maosong Sun. 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, page 4643–4652, New York, NY, USA. Association for Computing Machinery.

Yunyi Zhang, Jiaming Shen, Jingbo Shang, and Jiawei Han. 2020. Empower entity set expansion via language model probing. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8151–8160, Online. Association for Computational Linguistics.