MRHF: Multi-stage Retrieval and Hierarchical Fusion for Textbook Question Answering

Peide Zhu[1(✉)], Zhen Wang[2], Manabu Okumura[2], and Jie Yang[1]

[1] Delft University of Technology, Delft, Netherlands
{p.zhu-1,j.yang-3}@tudelft.nl
[2] Tokyo Institute of Technology, Tokyo, Japan
wzh@lr.pi.titech.ac.jp, oku@pi.titech.ac.jp

Abstract. Textbook question answering is challenging as it aims to automatically answer various questions on textbook lessons with long text and complex diagrams, requiring reasoning across modalities. In this work, we propose **MRHF**, a novel framework that incorporates dense passage re-ranking and the mixture-of-experts architecture for TQA. MRHF proposes a novel query augmentation method for diagram questions and then adopts multi-stage dense passage re-ranking with large pretrained retrievers for retrieving paragraph-level contexts. Then it employs a unified question solver to process different types of text questions. Considering the rich blobs and relation knowledge contained in diagrams, we propose to perform multimodal feature fusion over the retrieved context and the heterogeneous diagram features. Furthermore, we introduce the mixture-of-experts architecture to solve the diagram questions to learn from both the rich text context and the complex diagrams and mitigate the possible negative effects between features of the two modalities. We test the framework on the CK12-TQA benchmark dataset, and the results show that MRHF outperforms the state-of-the-art results in all types of questions. The ablation and case study also demonstrates the effectiveness of each component of the framework.

Keywords: Textbook Question Answering · Information Retrieval · Mixture-of-Experts

1 Introduction

The Textbook Question Answering (TQA) task [13] aims at automatically answering questions designed for multimodal textbook lesson materials. Unlike the text-based machine reading comprehension and visual question answering (VQA) tasks, where the context is text or image only, TQA aims to answer multiple types of multimodal scientific questions with scientific knowledge contained in both the text context and scientific diagrams. The requirement to answer multiple types of questions by understanding both the long context and complex diagrams makes TQA a challenging task.

P. Zhu, Z. Wang—Equal Contribution.

S. Rudinac et al. (Eds.): MMM 2024, LNCS 14555, pp. 98–111, 2024.
https://doi.org/10.1007/978-3-031-53308-2_8

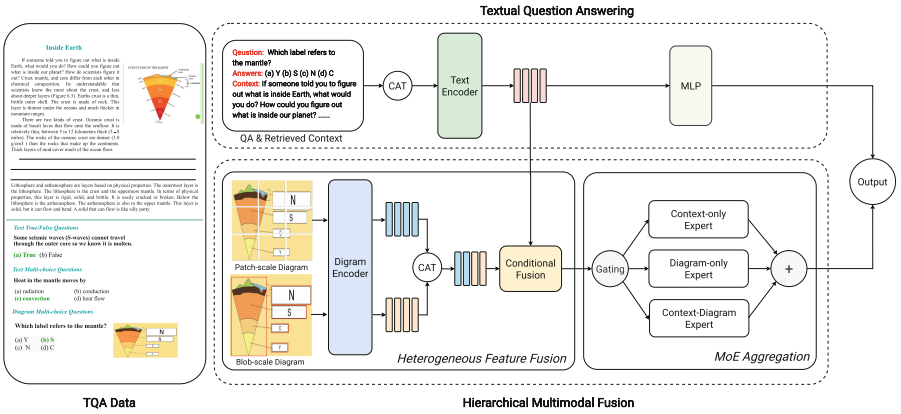


Fig. 1. The pipeline of our proposed MRHF. If a question does not contain a diagram, the upper textual question answering module will be activated to generate the answer directly. In contrast, if the question is a diagram question, the lower hierarchical multimodal fusion module will be activated.

The TQA task has attracted a lot of research efforts [3,4,14,16,17,19,20,29]. Despite previous progress, TQA remains a challenge. First, previous research retrieves several sentences from the whole corpus using the questions as queries. However, this approach overlooks two critical aspects. First, it fails to account for the fact that many questions are often related to sentences within the same paragraph, where a single sentence may not provide sufficient information for deducing the correct answer. Second, the method is inadequate for retrieving the correct context for diagram questions that tend to be vague and frequently depend on information within the diagrams, e.g., *which of the following labels is correct?*. Secondly, some diagram questions can be answered with the text context or diagram knowledge only, and features of the other modality may be negative for prediction [29]. However, previous research either ignores it or uses manually defined hyper-parameters as a solution, which are not adjustable and learnable for different instances.

In this work, we systematically address these challenges and propose **MRHF**, a novel framework for TQA with multi-stage context retrieval and hierarchical multimodal fusion (Fig. 1). To address the noisy context selection problem, we first apply the pretrained neural passage retriever for paragraph-level multi-stage context retrieval in the TQA task and show its effectiveness. We augment diagram questions with keywords extracted from texts associated with the question diagram and its related teaching diagrams. With the augmented queries, the retrieved paragraphs are more related to the question. Moreover, since the diagram questions can be related to certain specific regions of interest (RoI) or related to knowledge represented by the whole diagram, we propose a heterogeneous feature fusion (HFF) module to learn from different forms of diagram representation, including patch-level features extracted by visual-transformers

(ViT) and blob-level features extracted by YOLO. Furthermore, given that a question may relate exclusively to either the text context or the diagram, to suppress the negative effects from different modalities [29], we introduce the mixture-of-experts aggregation (MoEA). The MoEA consists of three experts, namely the context-only expert, the diagram-only expert, and the context-diagram expert; each is an MLP neural network following different encoding and fusion results, together with a trainable gating network which learns to give different weights to each expert to compose final prediction. Thus, the model is able to rely more on specific experts according to its features.

With MRHF, we perform extensive experiments on the CK12-TQA dataset and compare its performance with previous state-of-the-art (SOTA) methods. The experiment result shows that MRHF significantly surpasses previous methods in the TQA task on both text and diagram questions. We then conduct ablation studies and demonstrate the effectiveness of different components in MRHF. Our contributions in this paper can be summarized as follows: 1) we propose a multi-stage context retrieval method integrated with query augmentation and dense re-ranking, making the context we retrieved more relevant to questions; 2) we propose a hierarchical fusion method that includes the heterogeneous multimodal feature fusion and MoE, surpassing previous methods' performance on the diagram question; 3) detailed experiments and ablation studies prove the efficiency of different components in our method.

2 Related Work

As a complex multimodal QA task, TQA has attracted considerable research interest, particularly following the introduction of the (CK12-TQA) dataset [13]. Most efforts in TQA research can be categorized into three groups: context retrieval, diagram understanding, and question reasoning. Context retrieval is applied for gathering context knowledge related to the question. Most works extract sentences with lexical retrieval methods like TF-IDF [17], ElasticSearch [3], and Solr [4]. Some works like IGMN [16] propose to build essay-level contradiction entity-relationship graphs for reasoning in the long context. In addition to text lexical-based retrieval, ISAAQ [3], MoCA [29] further use different independent semantic-based methods for context retrieval. However, these retrieval methods suffer from noisy results because the sentence-level extraction cannot maintain enough information to answer the questions, and they lack the information from the diagram for retrieval. In recent years, dense passage retrieval methods based on pretrained models [11] and sentence transformers [22] have achieved great progress and shown competent performance in zero-shot retrieval scenarios. Our approach is the first to leverage the capabilities of dense retrievers for TQA.

Diagrams contain complex objects, text, and relations. A lot of attention has been put into effectively leveraging both text and diagram features for answering diagram questions. Some works explore the fine-grained relations among diagram components multimodal graphs for diagram QA, e.g., [13] translates the parsed

diagram graphs to factual sentences, [20] builds graphs for reasoning. Some other research leverages attention among the context text and the diagrams for diagram QA, e.g., [3] pretrain the diagram QA model on VQA datasets and leverage bottom-up and top-down attention for multimodal fusion, and [29] proposes to use patch-level diagram features generated by large pretrained visual transformers [1]. However, these methods cannot effectively leverage different-scale knowledge in the diagram, and there are possible negative effects between different modalities. Therefore, we introduce the Mixture-of-Experts (MoE) architecture to the TQA task. MoE has been proposed over two decades ago [7], designed to allow different sub-networks (experts) of a model to specialize for different samples with a learnable gating function. Different types of MoE have been proposed and applied to a range of tasks, including NLP and visual applications [2,24]. This is the first research that adopts MoE for the TQA task.

3 Method

3.1 The TQA Problem

Given a textbook QA dataset that consists of paragraphs $\mathcal{P} = \{P_1, P_2, \ldots, P_N\}$, a list of instructional diagrams $\mathcal{D} = \{d_1, d_2, \ldots, d_M\}$ and a list of questions $\mathcal{Q} = \{Q_1, Q_2, \ldots, Q_K\}$, where d_i denotes the i-th diagram and Q_i is the i-th question. A text question contains one question sentence q_i and its answer options A_i, where $A_i = \{a_{i,j}\}_{j=1}^{O}$ is the list of options. If Q_i is a multiple choice question, A_i is a list of O options, whereas it contains True or False if Q_i is a T/F question. If Q_i is a diagram question, then we represent it as $Q_i = \{q_i, A_i, \delta_i\}$ where δ_i is its corresponding question diagram. Then the answer inference of Q_i using a QA model with trainable parameters θ can be formulated as follows:

$$\hat{a}_i = arg \max_{a_{i,j} \in A_i} \Pr(a_{i,j} | q_i, C_i, [d_k, \delta_i]; \theta) \tag{1}$$

where $C_i \subset \mathcal{P}$ is the retrieved text context from the text contents and $d_k \in \mathcal{D}$ is the retrieved instructional diagram if Q_i is diagram question.

3.2 Multi-stage Context Retrieval

Although TQA lessons are extremely long (over 75% of them have at least 50 sentences), most (about 80%) questions require only several sentences from the same paragraph, and only some questions require information spread across the entire lesson. Instead of retrieving sentences like in previous work, we perform paragraph-level retrieval. We split paragraphs longer than a certain number (128, since most paragraphs are shorter than 128 words) of words into separate shorter paragraphs. In this way, the proposed method can gather cross-paragraph context and keep syntactic and semantic properties in each paragraph. Diagram questions pose distinct challenges for context retrieval that are neglected by previous research. Therefore we develop an extra query augmentation method for

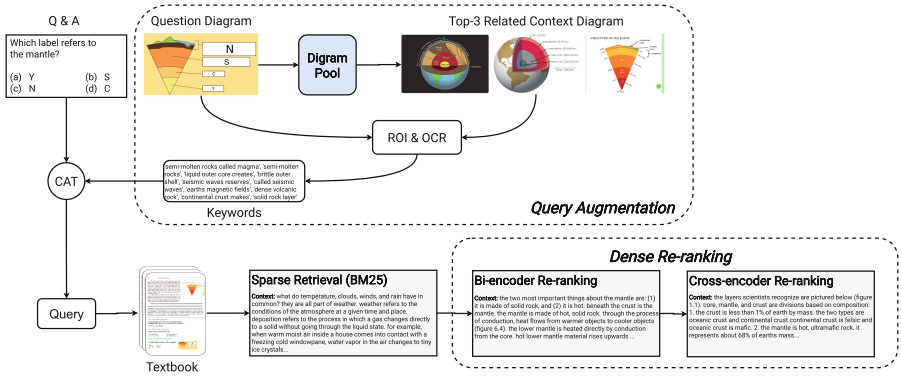


Fig. 2. Our multi-stage context retrieval pipeline includes query augmentation, sparse retrieval, and dense re-ranking.

diagram questions. We would introduce context retrieval pipelines for both diagram and text questions.

➢ **Query Augmentation for DQ.** The context knowledge of diagram questions includes both text and diagrams. As shown in Fig. 2, some diagram questions do not contain enough information to retrieve the text context that can answer it. The texts in the question diagram are replaced with option letters, which makes it difficult to leverage diagram annotations as extra vocabulary for retrieval. Therefore, instead of directly performing context retrieval with questions, we propose performing query augment first by using question diagrams as a bridge and retrieving related teaching diagrams as well as texts associated with them, and then performing context extraction using the same pipeline as text questions with the augmented query.

In detail, we first parse the diagrams that contain complex components like images, arrows, and text that convey critical information for understanding. To extract these components, we fine-tune YOLO(V5) [9] on AI2d [12] to recognize the positions and types of all these components. The AI2D dataset contains diagrams in the same style and annotations for diagrams, including each component's type and position information. Then we use OCR to recognize the text in each text block. Besides text contained in diagrams, we also extract extra text associated with them, including diagram captions, textbook sentences that reference them, and detailed introductions for teaching diagrams.

We then perform instructional diagram retrieval to find related diagrams in textbooks. As components representing the same concepts often have different colors in various diagrams, we first convert the diagram to grayscale to mitigate distractions introduced by colors. We then encode each diagram with the visual transformers model (ViT) and use the embedding of `[CLS]` token as its representation and rank the instructional diagrams according to the cosine similarity to the question diagram.

$$sim(d_j, \delta_i) = \cos(\mathrm{ViT}_{\texttt{[CLS]}}(d_j), \mathrm{ViT}_{\texttt{[CLS]}}(\delta_i)) \tag{2}$$

We choose top-3 instructional diagrams and combine all texts associated with them as associated text. Then we augment r_i with extracted keywords from the associated text. Then we apply the same context knowledge retrieval pipeline introduced in Sect. 3.2 with the augmented query.

➢ **Sparse Retrieval.** With the original QA pairs for text questions or augmented queries for diagram QA, we construct queries (r_i) for context retrieval by combining the questions and all answer options (except the *true*, *false*, *none*, and *all* options) with white spaces `[W]`:

$$r_i = q_i \texttt{[W]} a_{i,1} \texttt{[W]} a_{i,2} \texttt{[W]} \ldots a_{i,|A_i|}$$

and then perform multi-stage retrieval to obtain the QA context, as shown in Fig. 2. We apply **sparse retrieval** with BM25 [25], a well-adopted space vector-based probabilistic text retrieval method. In this step, we choose top-K_1 paragraphs (C_i^{BM25}) that have the most lexical similarity with the queries.

➢ **Dense Re-ranking.** Different from previous research that uses sparse models such as TF-IDF only, we use two-stage **dense re-ranking** to refine the context based on both lexical and semantic similarity to the question-answer pairs, as large pretrained language model-based dense passage retrieval has demonstrated substantial improvements in retrieval performance. In our dense re-ranker, we first employ a standard pretrained neural IR architecture [11] for a semantic **bi-encoder re-ranking** (BI). It uses the pretrained transformer encoder E_C which encodes the context paragraphs and E_Q on the queries into separate m-dimensional real-valued vectors and retrieves top$-K_2$ paragraphs C_i^{BI} in terms of cosine similarity:

$$sim(P_{i,j}, r_i) = \cos(E_C(P_{i,j}), E_Q(r_i)). \tag{3}$$

where $P_{i,j} \in C_i^{BM25}$. Then, we further leverage **cross-encoder re-ranking** (CE), which is used for matching text pairs by concatenating the query and target paragraph together, treating it as a sequence classification task, and performing full self-attention over the entire sequence. As reported in some research [23], the cross-encoder could have better performance than the bi-encoders with the sacrifice of efficiency. Therefore, we re-rank the semantic retrieval results C_i^{BI} with a pretrained cross-encoder.

$$sim(P_{i,j}, r_i) = \text{MLP}(E(\texttt{[CLS]} P_{i,j} \texttt{[SEP]} r_i \texttt{[SEP]})) \tag{4}$$

where $\text{MLP}(h_i) = W_2(W_1 h_i + b_1) + b_2$ is a multilayer perceptron network that takes the encoder E's output h_i as the input for calculating the final matching score, and W_1, W_2, b_1, b_2 are trainable parameters. We choose the top-3 paragraphs as the context passage.

3.3 Textual Question Answering

The process of choosing the correct answer to Text-MC questions is similar to answering the T/F questions, which can be interpreted as verifying whether

the context can support the claim of the question and the option pair. Inspired by [8], we transfer multiple-choice questions as a single-choice decision problem and treat T/F questions and Text-MC questions as a sequence bi-classification problem.

$$\begin{aligned} I_{i,j} &= \texttt{[CLS]}C_i\texttt{[SEP]}\hat{a}_{i,j}\texttt{[SEP]} \\ f^t_{i,j} &= \text{MLP}(\mathbf{E}_{\texttt{[CLS]}}(I_{i,j})) \end{aligned} \tag{5}$$

where $f_t = \Pr(\hat{a}_{i,j}|q_i, C_i)$ represents the predicted probability of the correctness of the j-th answer option $\hat{a}_{i,j}$ in A_i, and we use softmax to normalize the MLP's output of True/False probability to [0−1]. For Text-T/F questions and the option like *none* or *none of the above*, $\hat{a}_{i,j}$ is an empty string. For answer options such as *all* or *all of the above*, we concatenate all other options as the option text. To train the model, we label the sequence with the correct answer option as True, and others as False. We label all answer options as True when the correct answer is *all*. We use Cross-Entropy loss to train the model. For prediction, we chose the option of highest probability on True as the correct answer. To the questions with option *none*, we predict the correct answer is *none* when 1) the *none* option has the highest probability, or 2) all other options' probabilities are below 0.5. For the questions with option *all*, we predict the correct answer is *all* when 1) the *all* option has the highest probability, or 2) the probabilities of all other options are greater than 0.5. Similar to previous TQA researches [3,4,29], we perform pretraining with some extra datasets such as RACE [15] and SQuAD [21].

3.4 Hierarchical Multimodal Fusion

To answer the diagram questions, the solver should be able to reason over both the text (context, questions, and answer options) and the diagrams (question diagram and instructional diagrams). Many (40%) diagram questions require complex diagram parsing [3] and are relevant to certain regions of interest (RoI) or the relation and knowledge represented by the whole diagram. As pointed out in previous work, text contexts of a considerable proportion of diagram questions are rich enough to answer them, and features of the other modality may have negative effects. Therefore, we first propose a heterogeneous feature fusion module to learn from different contextualized diagram representations. Then we adopt the mixture-of-experts architecture to learn from different modality and their interaction.

➢ **Heterogeneous Feature Fusion (HFF).** To better leverage the features from the diagram, here we propose the HFF module. As mentioned above, we first parse the question diagram δ_i and get a list of blobs B_i. Then, we create the patch-level features V^p_i, and the blob-level features V^b_i of B_i using the ViT encoder, where $V^b_i = [\text{ViT}_{\texttt{[CLS]}}(B^j_i)]^{|B_i|}_0$. We concatenate them to generate a heterogeneous representation V^d_i of the diagram. For the j-th option $a_{i,j}$ of the question, we create the text features $V^t_{a_{i,j}}$ for the QA pair, and the text features $V^t_{c_{i,j}}$ for QA pair and the context with the trained text model. Instead of using all the features for further processing, here we use gated attention [30]

to find the most important features. The dual fusion of the text feature and the heterogeneous diagram features is calculated as follows:

$$\begin{aligned} U_{i,j} &= V^t_{c_{i,j}} W_u V^d_i \\ S_{i,j} &= \text{softmax}(U_{i,j}/\sqrt{d_{V^d_i}}) \\ Z_{i,j} &= W_s[V^t_{c_{i,j}} : S^T_{i,j} V^d_i] \\ z_{i,j} &= \text{MaxPooling}(Z_{i,j}) \\ v_i &= \text{MaxPooling}(V^t_{c_{i,j}}) \\ g_{i,j} &= \text{sigmoid}(W_g z_{i,j}) \\ h_{i,j} &= g_{i,j} \cdot \tanh(z_{i,j}) + (1 - g_{i,j}) \cdot v_i \end{aligned} \tag{6}$$

where W_u, W_s, W_g are trainable weights.

We then obtain the full-text ([question & answer options & context]) guided representation $h^c_{i,j}$ by using the equation. Similarly, we calculate the qa-text ([question & answer options]) guided representation $h^a_{i,j}$. To calculate the diagram guided representation, we substitute V^d_i and $V^t_{c_{i,j}}$ in the equation obtain $j^d_{i,j}$, and then obtain $j^a_{i,j}$ in the similar way.

➢ **Mixture-of-Experts Aggregation (MoEA).** Since diagram questions can be answered by using only context, or only diagram, or must leverage both diagram and context, we design different experts to handle different situations and use the mixture-of-experts to aggregate the results of those experts.

We first combine $h^c_{i,j}$ and $j^d_{i,j}$ to form $u^c_{i,j}$ representing the fusion of diagram and text with context, and also combine $h^a_{i,j}$ and $j^a_{i,j}$ to form $u^a_{i,j}$ representing the fusion of diagram and text without context, to input different experts in the next step. Based on the text features and the dual fusion representations, we design three question solvers, namely the text question solver $f^t_{i,j}$ which is MLP for $V^t_{c_{i,j}}$, the diagram-only solver $f^a_{i,j}$ which is MLP for $u^a_{i,j}$, and the context-diagram solver $f^c_{i,j}$ which is MLP for $u^c_{i,j}$. We utilize the MoEA with a learnable gating function G to automatically learn to put different weights on these different solvers (experts). We adopt the simple yet widely used gating function [10] to calculate the weights by multiplying a trainable matrix W_γ with the input and then normalize the weights by softmax. We concatenate $u^a_{i,j}$ and $u^c_{i,j}$ as input to the gating function.

$$\begin{aligned} \mu_{i,j} &= G(u^a_{i,j}, u^c_{i,j}) \\ G &= softmax(W_\gamma[u^a_{i,j} : u^c_{i,j}]) \\ f^{MoE}_{i,j} &= \mu_{i,j} \cdot [f^t_{i,j}, f^a_{i,j}, f^c_{i,j}] \end{aligned} \tag{7}$$

where the output of the gating function μ is a 3-dimension vector $[\mu_0, \mu_1, \mu_2]$ which represents weights for the two experts. The weighted sum of the outputs of the three experts ($f^{MoE}_{i,j}$) is the final prediction for the $j-$th answer option of diagram multiple-choice question i.

4 Experiment

4.1 Experimental Settings

➢ **Datasets.** We conduct the model evaluation on the CK12-TQA dataset, which contains textbook lessons, different types of questions, and rich textbook diagrams and has become the benchmark dataset for TQA research. Data samples in CK12-TQA can be categorized by the type of questions into three groups: true/false (T/F), text multiple choice (T-MC), and diagram multiple choice (D-MC) questions. The details of the datasets we use are shown in Table 1.

Table 1. Statistics of TQA dataset.

Dataset	Train	Dev	Test	Total	Options
CK12-TQA	15,154	5,309	5,797	26.260	–
–T/F	3,490	998	912	5,400	2
–T-MC	5,163	1,530	1,600	8,293	4–7
–D-MC	6,501	2,781	3,285	12,567	4

➢ **Implementation Details.** To create the dataset for pretraining, we perform named entity recognition and POS tagging using SpaCy [6]. For text context retrieval, we use the pretrained bi-encoders (MPNet [27]) and cross-encoders (MiniLM [28]) provided by Sentence-Transformers library [22]. We extract keywords from texts associated with retrieved instructional diagrams with RAKE [26] and use the top-5 extracted keywords for query augmentation. We use RoBERTa-large [18] for the sequence classification model. We train YOLO [9] on the AI2D dataset for 50 epochs for diagram parsing. For diagram retrieval and encoding, we use the visual transformers pretrained via masked autoencoders [5]. We finetune the model on the CK12-TQA dataset for 10 epochs on one NVIDIA A40 GPU with an initial learning rate at $1e^{-6}$, and the batch size is 4.

4.2 Experiment Results

We evaluate the proposed framework's performance in terms of its accuracy ($\frac{\#correct}{\#questions}$) on T/F, text, and diagram MC questions in both validation and test splits. We compare its performance with the previous state-of-the-art (SOTA) models, including single model approaches **IGMN** [16], **XTQA** [19] and **MHTQA** [4], as well as the ensemble approaches: **ISAAQ** [3] and **MoCA** [29]. The main results are shown in Table 2.

Table 2. Experimental results on the CK12-TQA validation and test splits in terms of accuracy. ⋆ means we choose the best single model from the ensemble solutions for comparison. We train these models with the context extracted using methods introduced in this paper.

Model	Val Set					Test Set				
	T/F	T-MC	T-All	D-MC	All	T/F	T-MC	T-All	D-MC	All
Random	50.86	23.66	34.40	25.83	29.91	50.37	22.93	32.89	24.80	28.31
Single Model										
IGMN	57.41	40.00	46.88	36.35	41.36	–	–	–	–	–
XTQA	58.24	30.33	41.32	32.05	36.46	56.22	33.40	41.67	33.34	36.95
ISAAQ-IR⋆	78.26	67.52	71.76	53.83	62.37	77.74	68.94	72.13	50.50	59.87
MHTQA	82.87	69.22	74.61	54.87	64.27	–	–	–	–	–
MoCA-IR⋆	–	73.33	–	54.15	–	–	–	–	52.12	–
MRHF	**87.48**	**76.80**	**81.01**	**56.27**	**67.90**	**86.51**	**79.19**	**81.85**	**53.97**	**66.05**
w/o CE	84.37	75.36	78.92	55.09	66.29	82.90	77.44	79.42	52.15	63.97
w/o BI&CE	85.47	74.84	79.04	50.38	63.88	82.24	78.25	79.70	51.08	63.48
Ensemble Model										
ISAAQ	81.36	71.11	75.16	55.12	64.66	78.83	72.06	74.52	51.81	61.65
MoCA	81.56	76.14	78.28	56.49	66.87	81.36	76.31	78.14	53.33	64.08
MRHF	**87.88**	**78.95**	**82.48**	**56.67**	**68.80**	**86.62**	**80.00**	**82.40**	**54.55**	**66.62**

We first compare our method with previous methods in a single-model setting. After further fine-tuning on the CK12-TQA dataset, **MRHF** achieves 81.01% in overall accuracy on the text questions of the validation set, and 81. 85% in the test set, which outperforms the previous best single-model MoCA by a margin of about 6.4% and 9.7% on validation and test sets separately in all text questions according to available data. MRHF also outperforms the SOTA single method on diagram multi-choice questions by a margin of 2.12% and 1.85% on the validation set and test set, respectively. Since previous SOTA methods ISAAQ and MoCA are both ensemble models which ensemble multiple models trained on different retrieved results, we also compare MRHF's performance in the ensemble-model setting. It can be found that the accuracy of ensemble MRHF is further improved and achieves new SOTA performance. Moreover, even the single model MRHF exceeds the ensemble MoCA model. These results demonstrate the effectiveness of our proposed MRHF framework.

5 Ablation Studies

5.1 Quantitative Analysis

➢ **Query Augmentation.** We first investigate the impact of query augmentation. Results are shown in Table 3. For the setting without text context, we replace the $I_{i,j}$ with answer option text. As the result shows, the performance

on the validation set deteriorates to 46.46% with about 9.8% decline and 3.04% decline on the test set, which first demonstrates the importance of context features. Then we remove the query augmentation, the performance of MRHF declines 1.6% and 1.46% on the validation and test set, respectively, which can demonstrate the effectiveness of query augmentation in context retrieval in Sect. 3.2.

Table 3. Ablations study on the impact of context text, query augment (AGM), and mixture-of-experts on the performance of answering diagram MC questions in CK12-TQA validation and test splits.

Split	MRHF	w/o Text	w/o AGM	w/o HFF	w/o MoEA
Val	**56.27**	46.46	54.67	55.34	54.95
Test	**53.97**	49.99	51.57	51.18	51.96

➢ **Dense Re-ranking.** We then finetune MRHF on CK12-TQA data with context extracted using different retrieval methods, and report MRHF-BM25, MRHF-BI, and MRHF-CE results in Table 2. First, we observe that with BM25 only, it outperforms previous SOTA models that utilize similar lexical retrieval methods, revealing the effectiveness of paragraph-level retrieval. Second, applying Bi-Encoder and Cross-Encoder Re-ranking can further improve the performance, demonstrating that applying the neural ranking models in a zero-shot setting on the TQA task is effective. Third, we observe that the model achieves over 53% accuracy on diagram questions with a text context only, showing the importance of text context in answering diagram questions.

➢ **Hierarchical Multimodal Fusion.** We examine the HFF and MoEA employed in our hierarchical multimodal fusion module, as demonstrated in Table 3. When we remove the HFF, the overall performance drops to 55.34% and 51.18% on the validation and test sets, which demonstrates the usefulness of HFF. We then eliminate the MoEA, and the performance decreases by 1.32% on the validation set and 2.01% on the test set, indicating the effectiveness of MoEA.

5.2 Case Studies

Fig. 3 illustrates the impacts of multi-stage context retrieval with query augmentation for diagram questions. We show the question diagrams with the bounding box of all blobs detected by the diagram parsing step. The example in the first row shows that without AGM, the retrieved context is about *food* topic, which totally drifts off the actual topic on *rain*. By contrast, with AGM, the multi-stage context retrieval method can retrieve the exact context paragraph. Examples in other rows show that AGM can still help improve the retrieval results even for diagram questions where there is rich text information. The last column reports

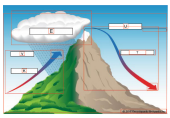
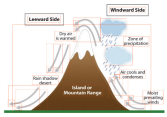
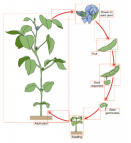
Question	Answer Options	Question Diagram	Keywords	w/o-AGM	AGM	Option Probs
Which label refers to rains?	(a) V (b) T (c) U (d) E		'wind moves particles suspension', '21 wind transports particles', 'oxidiz- ing agent', '23 sand dunes form', 'wind deposits sand', 'desert sand dunes', 'sand dunes line', 'sand dunes', 'ways depending', 'dust storm'	look at the nutrition facts label in figure 17.7. instructions at the right of the label tell you what to look for. at the top of the label, look for the serving size. the serving size tells you how much of the food you should eat to get the nutrients listed on the label. for this food, 1 cup is a serving. the calories in one serving are listed next. in this food, there are 250 calories per serving. next on the nutrition facts label, look for the percent daily values (% dv) of several nutrients.	the diagram is a representation of how a rain shadow is formed. a rain shadow is a dry region of land on the side of a mountain range that is protected from the prevailing winds. prevailing winds are the winds that occur most of the time in a particular location on the earth. the protected side of a mountain range is also called the lee side or the down-wind side. prevailing winds carry air toward the mountain range. as the air rises up over a mountain range, the air cools, water vapor condenses, and clouds form.	Ans, Tex, Dia, T&D (a), 0.087, 0.145, 0.102, 0.265 (b), 0.294, 0.325, 0.334, 0.172 (c), 0.122, 0.177, 0.148, 0.276 (d), 0.501, 0.354, 0.416, 0.287 W, 0.036, 0.048, 0.917
From the diagram, what happens after the presence of moist prevailing winds?	(a) rain shadow desert (b) air cools and condenses (c) dry air is warmed (d) zone of precipitation'		'wind moves particles suspension', '21 wind transports particles', 'physical models liquid', '23 sand dunes form', 'wind deposits sand', 'desert sand dunes', 'sand dunes line', 'sand dunes', 'physical model', 'ways depending'	this diagram shows the effect of rains on hills. the moist air from the green side of the hills rising up in the air and condenses as water vapor. this is called precipitation. the other side of the hill from the rain shadow region. rain shadow region is a region having little rainfall because it is sheltered from prevailing rain-bearing winds by a range of hills. the dry air descends from this region.	this diagram shows the effect of rains on hills. the moist air from the green side of the hills rising up in the air and condenses as water vapor. this is called precipitation. the other side of the hill from the rain shadow region. rain shadow region is a region having little rainfall because it is sheltered from prevailing rain-bearing winds by a range of hills. the dry air descends from this region.	Ans, Tex, Dia, T&D (a), 0.010, 0.247, 0.195, 0.017 (b), 0.681, 0.294, 0.423, 0.471 (c), 0.012, 0.218, 0.188, 0.020 (d), 0.297, 0.241, 0.195, 0.493 W, 0.094, 0.005, 0.902
Identify the part of the plant which functions as nourishment for the embryo, dispersal to a new location, and dormancy during unfavorable conditions.	(a) Seed (b) Fruit (c) Flower (d) Seedling		'tremendous feather displays', 'unfavorable mutations selected', 'sand dunes form', 'mutation creates variation', 'namib desert beetle', 'mutations reproduction', 'favorable mutations', 'sand dune', '11 genetic mutation', 'biological diversity'	the diagram shows the internal structure of a corn kernel or corn seed. the seed has three main parts seed coat, endosperm and the embryo. the seed coat is the outermost part of the seed and it protects both the endosperm and the embryo. the endosperm is below the seed coat. it stores food in the seed. the embryo is what gives rise to a new plant. early growth and development of a plant embryo inside a seed is called germination. the embryo consists of the embryonic leaves, cotyledon and the primary root.	..... a seed is a reproductive structure that contains an embryo and a food supply, called endosperm. both the embryo and endosperm are enclosed within a tough outer coating, called a hull (or shell). you can see these parts of a seed in figure 10.14. an embryo is a zygote that has already started to develop and grow. early growth and development of a plant embryo inside a seed is called germination. the seed protects and nourishes the embryo and gives it a huge head start in the "race" of life. both a parent plant and its offspring are better off if they dont grow too closely together.....	Ans, Tex, Dia, T&D (a), 0.390, 0.251, 0.380, 0.201 (b), 0.290, 0.255, 0.218, 0.370 (c), 0.114, 0.256, 0.210, 0.163 (d), 0.206, 0.238, 0.192, 0.266 W, 0.071, 0.004, 0.924

Fig. 3. Case study of AGM and MoEA from test set. Under "Option Probs", "Ans" means answer candidates, and red is the correct answer, "Tex" means context-only expert, "Dia" means diagram-only expert, "T&D" means context-diagram expert, "W" means the weight of different experts.

the prediction by each expert as well as the learned weights for them. The results first suggest that the context-diagram expert has the largest weight. Second, through the mixture, although the single context-diagram expert makes a wrong prediction, the overall results adjusted by text and diagram-only experts successfully choose the correct answer.

6 Conclusion

In this paper, we propose a concise framework MRHF to address the challenges in the textbook question answering task, especially for diagram-related QA pairs. Experiment on CK12-TQA shows that our proposed framework can effectively solve the TQA problem, outperforming previous SOTA results on all types of questions. Even single-model MRHF can achieve considerable performance compared to previous ensemble models, which can significantly simplify the training, maintenance, and deployment of the TQA systems. Ablation studies further demonstrate the effectiveness of its components, including multi-stage context retrieval with query augmentation for diagram questions, multimodal conditional fusion, and the mixture-of-experts architecture. In the future, we will further study and unravel the challenges in multimodal question answering.

References

1. Dosovitskiy, A., et al.: An image is worth 16 × 16 words: transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
2. Fedus, W., Dean, J., Zoph, B.: A review of sparse expert models in deep learning. arXiv preprint arXiv:2209.01667 (2022)
3. Gómez-Pérez, J.M., Ortega, R.: ISAAQ-mastering textbook questions with pre-trained transformers and bottom-up and top-down attention. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 5469–5479 (2020)
4. He, J., Fu, X., Long, Z., Wang, S., Liang, C., Lin, H.: Textbook question answering with multi-type question learning and contextualized diagram representation. In: Farkaš, I., Masulli, P., Otte, S., Wermter, S. (eds.) ICANN 2021. LNCS, vol. 12894, pp. 86–98. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-86380-7_8
5. He, K., Chen, X., Xie, S., Li, Y., Dollár, P., Girshick, R.: Masked autoencoders are scalable vision learners. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 16000–16009 (2022)
6. Honnibal, M., Montani, I.: spaCy 2: natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing (2017). to appear
7. Jacobs, R.A., Jordan, M.I., Nowlan, S.J., Hinton, G.E.: Adaptive mixtures of local experts. Neural Comput. **3**(1), 79–87 (1991)
8. Jiang, Y., et al.: Improving machine reading comprehension with single-choice decision and transfer learning. arXiv preprint arXiv:2011.03292 (2020)
9. Jocher, G., et al.: ultralytics/yolov5: v4.0 - nn.SiLU() activations, Weights & Biases logging, PyTorch Hub integration. Zenodo, January 2021. https://doi.org/10.5281/zenodo.4418161
10. Jordan, M.I., Jacobs, R.A.: Hierarchical mixtures of experts and the EM algorithm. Neural Comput. **6**(2), 181–214 (1994)
11. Karpukhin, V., et al.: Dense passage retrieval for open-domain question answering. arXiv preprint arXiv:2004.04906 (2020)
12. Kembhavi, A., Salvato, M., Kolve, E., Seo, M., Hajishirzi, H., Farhadi, A.: A diagram is worth a dozen images. In: Leibe, B., Matas, J., Sebe, N., Welling, M. (eds.) ECCV 2016. LNCS, vol. 9908, pp. 235–251. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46493-0_15
13. Kembhavi, A., Seo, M., Schwenk, D., Choi, J., Farhadi, A., Hajishirzi, H.: Are you smarter than a sixth grader? Textbook question answering for multimodal machine comprehension. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 4999–5007 (2017)
14. Kim, D., Kim, S., Kwak, N.: Textbook question answering with multi-modal context graph understanding and self-supervised open-set comprehension. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 3568–3584 (2019)
15. Lai, G., Xie, Q., Liu, H., Yang, Y., Hovy, E.: RACE: large-scale reading comprehension dataset from examinations. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pp. 785–794 (2017)
16. Li, J., Su, H., Zhu, J., Wang, S., Zhang, B.: Textbook question answering under instructor guidance with memory networks. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 3655–3663 (2018)

17. Li, J., Su, H., Zhu, J., Zhang, B.: Essay-anchor attentive multi-modal bilinear pooling for textbook question answering. In: 2018 IEEE International Conference on Multimedia and Expo (ICME), pp. 1–6. IEEE (2018)
18. Liu, Y., et al.: Roberta: a robustly optimized Bert pretraining approach. arXiv preprint arXiv:1907.11692 (2019)
19. Ma, J., Chai, Q., Liu, J., Yin, Q., Wang, P., Zheng, Q.: XTQA: span-level explanations for textbook question answering (2023)
20. Ma, J., Liu, J., Wang, Y., Li, J., Liu, T.: Relation-aware fine-grained reasoning network for textbook question answering. IEEE Transactions on Neural Networks and Learning Systems (2021)
21. Rajpurkar, P., Zhang, J., Lopyrev, K., Liang, P.: Squad: 100,000+ questions for machine comprehension of text. arXiv preprint arXiv:1606.05250 (2016)
22. Reimers, N., Gurevych, I.: Sentence-BERT: sentence embeddings using Siamese BERT-Networks. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics, November 2019
23. Reimers, N., Gurevych, I.: Sentence-BERT: sentence embeddings using Siamese BERT-Networks. arXiv preprint arXiv:1908.10084 (2019)
24. Riquelme, C., et al.: Scaling vision with sparse mixture of experts. Adv. Neural Inf. Process. Syst. **34**, 8583–8595 (2021)
25. Robertson, S.E., Walker, S.: Some simple effective approximations to the 2-Poisson model for probabilistic weighted retrieval. In: Croft, B.W., van Rijsbergen, C.J. (eds.) SIGIR '94, pp. 232–241. Springer, London (1994). https://doi.org/10.1007/978-1-4471-2099-5_24
26. Rose, S., Engel, D., Cramer, N., Cowley, W.: Automatic keyword extraction from individual documents. Text Min. Appl. Theory **1**(1–20), 10–1002 (2010)
27. Song, K., Tan, X., Qin, T., Lu, J., Liu, T.Y.: MPNet: masked and permuted pre-training for language understanding. Adv. Neural Inf. Process. Syst. **33**, 16857–16867 (2020)
28. Wang, W., Wei, F., Dong, L., Bao, H., Yang, N., Zhou, M.: MINILM: deep self-attention distillation for task-agnostic compression of pre-trained transformers. Adv. Neural Inf. Process. Syst. **33**, 5776–5788 (2020)
29. Xu, F., et al.: MoCA: incorporating domain pretraining and cross attention for textbook question answering. Pattern Recognit. **140**, 109588 (2023)
30. Zhao, Y., Ni, X., Ding, Y., Ke, Q.: Paragraph-level neural question generation with maxout pointer and gated self-attention networks. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 3901–3910 (2018)