PROGRESSIVE EVIDENCE REFINEMENT FOR OPEN-DOMAIN MULTIMODAL RETRIEVAL QUESTION ANSWERING

Shuwen Yang*, Anran Wu*, Xingjiao Wu[†], Luwei Xiao*, Tianlong Ma*, Cheng Jin[†], Liang He*

*East China Normal University, Shanghai, China †Fudan University, Shanghai, China

ABSTRACT

Pre-trained multimodal models have achieved significant success in retrieval-based question answering. However, current multimodal retrieval question-answering models face two main challenges. Firstly, utilizing compressed evidence features as input to the model results in the loss of finegrained information within the evidence. Secondly, a gap exists between the feature extraction of evidence and the question, which hinders the model from effectively extracting critical features from the evidence based on the given question. We propose a two-stage framework for evidence retrieval and question-answering to alleviate these issues. First and foremost, we propose a progressive evidence refinement strategy for selecting crucial evidence. This strategy employs an iterative evidence retrieval approach to uncover the logical sequence among the evidence pieces. It incorporates two rounds of filtering to optimize the solution space, thus further ensuring temporal efficiency. Subsequently, we introduce a semi-supervised contrastive learning training strategy based on negative samples to expand the scope of the question domain, allowing for a more thorough exploration of latent knowledge within known samples. Finally, in order to mitigate the loss of fine-grained information, we devise a multi-turn retrieval and question-answering strategy to handle multimodal inputs. This strategy involves incorporating multimodal evidence directly into the model as part of the historical dialogue and question. Meanwhile, we leverage a cross-modal attention mechanism to capture the underlying connections between the evidence and the question, and the answer is generated through a decoding generation approach. We validate the model's effectiveness through extensive experiments, achieving outstanding performance on WebQA and MultimodelQA benchmark tests.

Index Terms— Web Question Answering, Multimodal Retrival, Transformer.

Question: In the photo of the woman who sings "Why You Gotta Be So Mean," what is the red shape next to her?

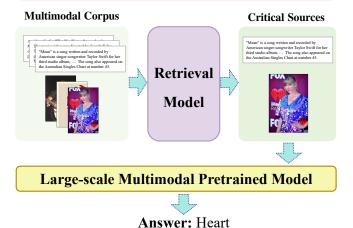


Fig. 1: Multimodal Retrieval Question-Answering Overall Workflow. The task requires the system to first identify critical sources from candidate sources through the retrieval system, and then use these key sources as clues to infer the answers to the questions.

1. INTRODUCTION

With the continuous advancement of World Wide Web (WWW) technologies, people's activities on the Internet have expanded, resulting in an exponential growth of data generated every day. Due to the sheer volume of information [1], individuals struggle to effectively process and sift through this data, leading to the risk of information overload. Furthermore, contemporary internet data is no longer limited to pure text but includes richer multimodal elements such as images, text, and tables. Consequently, extracting crucial information from this vast multimodal information has become a significant challenge [2–5], enabling better

Shuwen Yang and Anran Wu contributed equally to this work. Corresponding author: Xingjiao Wu (e-mail: xjwu_cs@fudan.edu.cn).

management and utilization of this extensive information resource. Quality, credibility, and accessibility of information have also become focal points to ensure that users can obtain valuable and accurate content from the massive amount of information available. Through persistent efforts, researchers have explored technical approaches using multimodal retrieval question-answering methods to extract key evidence from given multimodal composite data sources and answer questions [6, 7].

Multimodal retrieval-based question answering involves reasoning by identifying key evidence. However, in practice, the number of candidate evidence is substantial, and these critical pieces of evidence have strict sequential dependencies [8]. This leads to exceedingly high time and space complexity in the reasoning process. To effectively address this challenge, existing approaches [6, 8, 9] first employ encoders to extract features from the input evidence. Subsequently, pooling operations are utilized to reduce the dimensionality of features. Finally, the reduced-dimensional evidence features are employed for reasoning, and a generation mechanism is employed to generate key evidence and answers. Nevertheless, this approach overlooks specific issues. Firstly, dimensionality reduction results in information loss. Additionally, due to the dimensionality reduction and independently performed feature extraction, a gap exists between the evidence and the original question during feature extraction, leading to a disconnection between the key features extracted and the question.

To mitigate the loss of evidence information, researchers [10–12] have proposed employing a classification-based approach to traverse the matching of questions and key evidence. The specific methodology involves utilizing evidence and question encoders to extract features from the evidence and question. Subsequently, the cosine similarity between the question and the features is calculated or integrated. Finally, a classifier is employed to determine whether a correlation exists between individual questions and individual items of evidence. This approach addresses the information loss caused by dimensionality reduction. However, the exhaustive and simplistic classification approach ignores the logical relationships among the evidence.

Through analysis, we have discovered that finding a more optimal approach to maximize the identification of effective evidence and address the curse of dimensionality constitutes a central challenge in multimodal retrieval-based question answering. It is evident that the more correct key evidence we find and the more effectively we connect them, the more accurate the obtained answers will be. Therefore, an intuitive idea arises: Can we eliminate as much irrelevant information as possible? Consequently, we propose a stepwise evidence refinement retrieval strategy comprising two steps: initial screening and iterative retrieval. The initial screening leverages the calculation of distances between question features and candidate features to identify potential candidate

evidence while excluding spurious samples that are entirely irrelevant to the question. To enhance the prediction of matching scores between questions and evidence, we employ contrastive learning [13, 14] to train the question encoder and evidence encoder. The iterative retrieval step utilizes the question and the evidence identified during the initial screening to further predict key evidence.

However, when applying contrastive learning for initial screening, traditional contrastive learning training often only utilizes key evidence and questions for comparison. The training effectiveness can be severely impacted, especially when there is a significant amount of interfering evidence in the dataset. We propose a negative sample semi-supervised contrastive learning training strategy to overcome this challenge. By reconstructing the question pool, we fully leverage interfering evidence to train the model, effectively filtering out interfering evidence and enhancing the robustness of the initial screening model.

During the question-answering stage, the model needs to reason the answer based on the question and the retrieved key evidence as input. Currently, the general approach for answering questions based on retrieved evidence is directly extracting features from the retrieved multimodal information and then performing question-answering. However, feature extraction results in the loss of some fine-grained details, which hinders effective reasoning based on these details. Therefore, a more favorable approach to address this challenge would be directly inputting the multimodal corpus into the model for question-answering without losing fine-grained information. Inspired by recent works [15, 16], we expect the model to directly process the initial information from the input sources rather than the pooled features of the sources, thereby mitigating the loss of fine-grained information. Thus, we propose a large-scale multimodal processing model that serves the question-answering task with a focus on key evidence. To further ensure the logical coherence among different modalities of evidence, we introduce an innovative multi-turn dialogue mechanism during model training. Through this mechanism, the model can interactively incorporate the evidence as dialogue history with the final input question and capture the intermodal relationships among the evidence using crossattention mechanisms, allowing multiple images to participate in the answer reasoning process simultaneously.

To validate the efficacy of the proposed method, we conducted extensive experiments on two widely used multisource visual question-answering datasets WebQA [6] and MultiModelQA [7]. Our approach achieved state-of-the-art (SOTA) performance on both the retrieval and question-answering tasks of the WebQA and MultiModelQA datasets, verifying its effectiveness and highlighting its superiority over existing methods.

In a nutshell, our contributions can be summarized as follows:

• In order to minimize information loss and fully ex-

ploit the logical reasoning relationships among the evidence, we propose a stepwise evidence refinement retrieval strategy. This strategy leverages the inherent logical connections among the evidence to effectively constrain the solution space, thereby enhancing the selection of key evidence. Additionally, our approach also achieves an efficient reduction in space complexity.

- We propose a negative sample semi-supervised contrastive learning training strategy to address the issue of disregarding interfering samples during the initial screening of contrastive learning. This scheme resolves the problem of underutilization of interfering samples during training, enhancing the robustness of the model and improving the accuracy of the initial screening process.
- To further ensure the logical coherence among different modalities of evidence, we propose a novel multiturn dialogue visual question-answering approach for handling the task of visual question answering with multiple-source evidence inputs. This approach mitigates the disconnection between evidence feature extraction and question processing, effectively preserving the latent information among the evidence and significantly boosting question-answering performance.
- We conduct comprehensive experiments and detailed analysis on publicly available datasets WebQA and MultiModelQA. Our method achieves state-of-the-art (SOTA) results on both the retrieval and questionanswering tasks, validating its significance.

2. RELATED WORK

Information Retrieval-Based Question and Answering. Paragraph retrieval has consistently been a vital component of information retrieval for question-answering tasks [17]. This task aims to utilize neural networks to locate relevant paragraphs within a corpus, using these paragraphs and questions as context for reasoning and generating answers. Strong sparse vector space models such as TF-IDF or BM25 have traditionally served as standard methods extensively applied to various QA tasks [18-21]. Recent research has also explored enhancing text-based retrieval by utilizing external structured information, such as knowledge graphs and Wikipedia hyperlinks [22, 23]. Existing retrieval-based QA methods can be classified into two categories: cross-attention models and dual-attention models. Cross-attention models [19,24–27] initially establish connections between queries and documents, then employ a single encoder equipped with cross-attention mechanisms to learn the potential interactions between queries and documents. Typically, cross-attention

models [11, 28–32] can achieve significant retrieval performance improvements but come at a higher computational cost. Dual-attention models, on the other hand, first encode queries and documents separately using distinct encoders and then employ dual-attention mechanisms to facilitate query-document matching.

Multimodal Retrieval-Based Question and Answering. Multimodal Multi-Hop Question Answering (MMQA) is similar to traditional retrieval-based question-answering tasks [6], but it differs in that, in MMQA, each question requires retrieving key clues from various multimodal information sources to reason and generate an answer. Additionally, the MMQA task introduces distractor items to evaluate a model's ability to retrieve key evidence and resist interference. Recently, there have been benchmark datasets [6, 7, 33] specifically designed for tackling complex questions that involve multimodal input and contextual reasoning.

Currently, most MMQA methods [10-12] typically encode source data and employ classification layers or maximum inner-product search to select evidence, which is then passed to a decoder for answer generation. However, these methods primarily filter evidence by predicting the relevance between the question and the evidence, making it challenging to identify difficult-to-spot evidence that is less directly related to the question but beneficial for inference. Furthermore, these methods often support single-image input only. Another category of methods [6, 8, 9] employs a different strategy by encoding source data into low-dimensional features and feeding these features along with the question to the decoder to predict evidence and generate answers. However, this approach results in significant information loss when compressing evidence, thereby limiting its questionanswering performance.

To address these issues, our approach adopts a two-stage evidence retrieval and answer generation process to avoid information loss. Moreover, we design a stepwise refined evidence retrieval strategy to tackle the challenge of identifying difficult key evidence. Additionally, to address the issue of multiple-image input, we introduce a multi-turn dialogue visual-language model to simultaneously process content from multiple images in a multi-turn dialogue manner, thus enhancing the model's performance.

3. APPROACH

This task is divided into two stages. First, given a question q and a set of candidate sources $E = \{e_0, e_1, ..., e_n\}$, where n is the number of candidate sources, which can include images, text, or table information. The task in the first stage is to retrieve the true sources $R = \{r_0, r_1, ..., r_m\}$ from the candidate sources E, where m is the number of true sources related to the answer. Then, in the second stage, the answering model needs to generate an answer based on the question q and the selected true sources R as context.

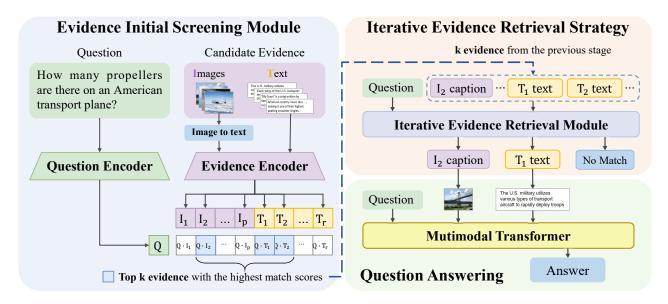


Fig. 2: Overall structure diagram of the PERQA model. It is mainly divided into two stages, one is the evidence retrieval stage which gradually refines the evidence, and the other is the question and answer stage based on multi-source input.

As depicted in Figure 2, our method, namely Progressive Evidence Refinement Question&Answering (PERQA), consists of two main steps. The first step involves evidence retrieval based on a step-wise evidence refinement strategy, which is used to precisely identify the true sources. The second step is the multi-cue question-answering stage, in which these two stages separately accomplish the tasks of evidence retrieval and answer generation.

3.1. Evidence Retrieval Stage Based on Progressive Evidence Refinement

The accuracy of retrieval in the context of a multi-hop retrieval question-answering method directly impacts the final accuracy of the answering process, making this step crucial. This stage is primarily divided into two steps: the first step involves the use of an **Evidence Initial Screening Module** (**EISM**) to filter out sources relevant to the question. We employ two simple question and evidence encoders to encode the question and evidence sources separately. We calculate the cosine similarity between the question and candidate source features as a matching score to filter potential genuine sources, thus narrowing down the retrieval scope for the next step. In the second step, we utilize an **Iterative Evidence Retrieval Strategy** (**IER**) to determine the true genuine sources from the sources filtered in the previous step.

3.1.1. Evidence Initial Screening Module (EISM)

Drawing inspiration from other text-matching methods, the Evidence Initial Screening Module comprises both a question encoder and an evidence encoder, both of which are language transformer models. However, for image sources, that cannot be directly processed by language transformers, we employ an approach that converts images into relevant textual information. In comparison to using CNN [34] or ViT [35] directly, this method incurs lower training costs and allows us to overlook the semantic gap between images and text during training. In our work, the image-to-text conversion methods we adopt include image captioning [36] and object detection [37], summarizing the entire image through overall descriptions and object descriptions. It's worth noting that we exclusively utilize the image-to-text approach during the retrieval stage. In contrast, in the question-answering stage, we directly employ the initial image information to accomplish the answering task.

We assume that the text representation of the i-th source is denoted as S_i . We start by inputting the question into the question encoder, obtaining the question feature f_Q , and then feeding each source into the evidence encoder, yielding source features f_{S_i} . Subsequently, we compute the similarity between the question feature and the candidate source features using cosine similarity, which serves as the matching score between the question and the candidate source. The formula is as follows:

$$f_Q = Q_{encoder}(Q) \tag{1}$$

$$f_{S_i} = E_{encoder}(S_i) \tag{2}$$

$$P_{cl}(Q, S_i) = \frac{f_Q \cdot f_{S_i}}{\|f_O\| \|f_{S_i}\|}$$
(3)

where $Q_{encoder}$ represents the question encoder, and $E_{encoder}$ represents the evidence encoder. After calculating

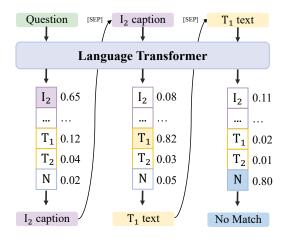


Fig. 3: The overall process of Iterative Evidence Retrieval Strategy.

the matching scores between the questions and each candidate source, we select the top-k candidate sources with the highest matching scores for further refinement in the next step.

3.1.2. Iterative Evidence Retrieval Strategy (IER)

In BERT-based question-answering methods [12, 38], to enhance inference and capture the relevance between questions and paragraphs, it is common to concatenate the question and paragraph and then input them into BERT to predict results. Our approach follows a similar pattern by concatenating the question and evidence and inputting them into BERT to predict their matching scores. This allows us to compute the most relevant evidence for a given question. As each question may have multiple key sources, we can iteratively append sources one after another to the question and the already retrieved sources to determine the matching score for the next source. Using this method, the model can capture implicit relationships between questions and sources, as well as between different sources, uncovering logical ordering relationships among sources, and ultimately enhancing the accuracy of retrieving subsequent sources.

Given a question Q and a set of candidate sources S, assuming that we have already determined the set of key sources R and the remaining set of candidate sources S-R, we calculate the matching score for the next source using the following method:

$$P_{ie}(Q, R, e) = reg(BERT(Q; R; e)) \tag{4}$$

Here, reg represents a regression module implemented using a linear layer, which is employed to compute the matching score. e is a candidate source such that $e \in S - R$. It is important to note that this process is iterative. Initially, R is an empty set, and S represents the coarse screening results for

the first round. Each time a key source is selected, we traverse the set of candidate sources and calculate the matching score. We select the source with the highest score to add to the set of critical sources and initiate the next round of retrieval until a termination signal is reached. In Figure 4, we begin by traversing and selecting the highest-scoring candidate evidence, I_2 , as the first key evidence. We then concatenate I_2 's caption with the question and continue to search for the following key evidence. Similarly, we choose the highest-scoring T_1 as the second key evidence, and so on, until we select a termination symbol. Typically, we include the termination symbol in the initial candidate evidence set. Due to the initial round of filtering, the candidate source set is not particularly large, ensuring time efficiency.

3.2. A question-answering model based on multi-sources input.

In multi-modal models [15, 16] that support multi-turn dialogues, images, and text can be linked as a history of conversations by concatenating the features of all images and text. A cross-attention mechanism is then used to capture relationships among these elements. Leveraging this, we utilize such a multimodal model for multi-turn conversations, inputting evidence as part of the conversation history and the question simultaneously. This approach allows us to generate answers. Since we've already acquired the most relevant evidence in the previous stage, we only need to select the most relevant pieces of evidence from the candidate set, significantly reducing computational costs.

As shown in Figure 5, our question-answering model consists of ViT [35] + LLaMA [39] + LORA [40] (Low-Rank Adaptation). For a higher-resolution model image, please refer to the appendix. We directly encode image information using ViT to obtain image features and encode text and question features using the LLAMA encoder. In intermediate layers, cross-attention is employed between image features and text features to capture relationships among different pieces of evidence and between evidence and questions. Finally, the LLAMA decoder decodes the combined features to obtain the final answer. Due to the large number of parameters in ViT and LLAMA, we freeze the parameters of the ViT and LLAMA models during training and only fine-tune the parameters of the LORA part.

3.3. Training

The training process consists of three steps. First, we train the two encoders used in the first step of evidence retrieval using a semi-supervised contrastive learning method with negative samples. Then, we separately train the iterative evidence retrieval model and the question-answering model using binary cross-entropy and cross-entropy loss functions.

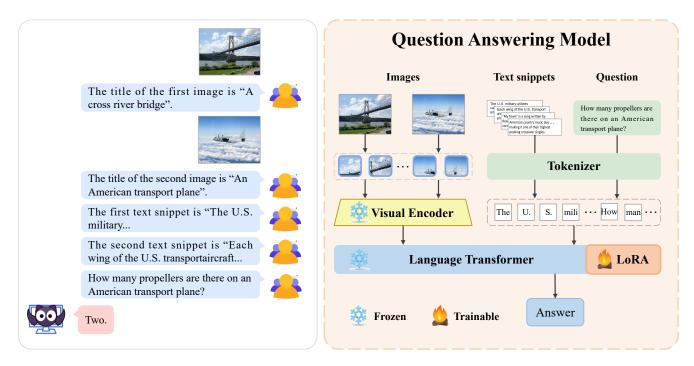


Fig. 4: The overall structure of the QA module. Our QA module uses evidence information as a historical dialogue to realize information interaction between evidence.

3.3.1. Negative sample semi-supervised contrastive learning (NSCL)

During this training phase, we followed the CLIP [41] approach, sampling multiple questions from the training set, each accompanied by a key piece of evidence, and organizing them into a batch for training. We use two separate language models to encode the information from both questions and evidence. The similarity between features is computed using cosine similarity as a measure of the match between the questions and evidence. During training, our goal is to maximize the similarity between the features of each question and its corresponding evidence while minimizing the similarity between the features of other evidence in the same batch. The objective function can be written as follows:

$$\mathcal{L}_{cl} = -\frac{1}{B} \sum_{i=1}^{B} \log \frac{exp(P_{cl}(Q_i, S_i))}{\sum_{j=1}^{B} exp(P_{cl}(Q_i, S_j))}$$
(5)

where B represents the batch size, and within the same remaining key samples are the ones we aim to predict as key batch, it contains multiple question-evidence pairs: $\{(Q_0, S_0), (Q_{\text{ENSO}}), (Q_{\text{ENSO}}), (Q_{\text{ENSO}})\}$. We want the model to infer the next key evidence amples. We want the model to infer the next key evidence

It can be observed that the naive contrastive learning training strategy requires each piece of evidence to be paired with a related question, but the dataset does not provide such related questions for distractor evidence. To address this issue, during the training phase, we treat each distractor evidence as its own related question and add

it to the batch for training. Formally, for each question Q_i , we sample a distractor source $S_{Q_i}^-$ associated with that question and include it in the batch. In this case, a batch of question-evidence pairs is updated as follows: $\{(Q_1,S_1),(S_{Q_1}^-,S_{Q_1}^-),...,(Q_B,S_B),(S_{Q_B}^-,S_{Q_B}^-)\}.$

Since the distractor sources for each question may be positively correlated with each other, and contrastive learning does not allow positively correlated samples, we sample only one distractor evidence per question to be added to the batch.

3.3.2. Iterative Evidence Retrieval Model Training

During the training phase of the iterative evidence retrieval model, we sample sets of question-key evidence pairs from the dataset. We divide the key evidence set into two parts: the first part serves as already retrieved key evidence, and the remaining key samples are the ones we aim to predict as key e_i while the distractor evidence serves as negative examples. We want the model to infer the next key evidence based on the question and the already retrieved key evidence, aiming for high matching scores with the remaining key evidence. Formally, assuming the question is Q, the retrieved evidence is R, the remaining key evidence is G, and the distractor evidence set is N, we use binary loss as the objective function:

$$\mathcal{L}_{ir} = -\left(\sum_{e^+ \in G} \log p(e^+) - \sum_{e^- \in N} \log p(e^-)\right)$$
 (6)

$$p(e) = P_{ie}(Q, R, e) \tag{7}$$

3.3.3. Question and Answering Model Training

For the training of the question and answering model, we employ the commonly used negative log-likelihood as the objective function for text generation. Formally, assuming the question is denoted as \mathbf{Q} and the retrieved evidence is denoted as R, the objective function for the generation part is defined as:

$$\mathcal{L}_g = -\sum_{i=1}^{l} -\log P_g(a_i|R;Q;a_{< i})$$
 (8)

Here, l represents the length of answer tokens, and P_g represents the probability score predicted by the question-answering model for the occurrence of a token.

4. EXPERIMENTS

4.1. Dataset

We conducted experiments on two of the most representative MMQA datasets: MultimodalQA [7] and WebQA [6]. Table 1 presents the statistical information for both datasets.

MultimodalQA [7] This dataset comprises multimodal questions from various modalities, manually annotated, including images, text, and tables. The questions in this dataset are generated from templates, with 16 question types and 13 requiring cross-modal retrieval and inference. Since test labels for this dataset have not been released, we report results solely on the validation set. Answers in MultimodalQA typically consist of phrases, and the evaluation metrics employed are Exact Match (EM) and Average F1.

WebQA [6] This dataset contains multi-hop, multimodal question-answer pairs, where each query requires 1-2 images or 1-2 text snippets to answer. The answers in WebQA are in free-form sentences. There are two evaluation metrics used: one for assessing the model's retrieval accuracy, which is the F1 metric, and another for evaluating the quality of the model's answer generation, which combines QA-FL and QA-Acc based on BARTScore [42]. QA-FL measures the fluency (grammatical and semantic coherence) between the generated answer and the reference, while QA-Acc assesses the overlap of key entities between the output answer and the reference.

4.2. Baselines

For our comparative experiments, we selected state-of-the-art (SOTA) baseline models from WebQA and MultimodalQA:

AutoRoute [7], ImplicitDec [7], VLP [6], VLP+VinVL [6], MuRAG [9], and SKURG [8]. AutoRoute [7] identifies the question modality and directs questions and input sources to the corresponding QA modules (textQ, tableQ, or imageQ) by employing a question-type classifier and utilizing different submodels to extract answers. This approach employs RoBERTa-large [43] for question type classification and textrelated question answering and uses VILBERT-MT [44] for visual-related question answering tasks. ImplicitDec [7] follows a similar method to AutoRoute but introduces a sequential concept. VLP [6] and VLP+VinVL [6] are encoder-decoder models based on transformers. They first extract pooled features of individual modal evidence using pretrained image encoders or text encoders and then concatenate these evidence features with the question as input to predict key evidence and generate answers. MuRAG [9] precomputes the encoding of candidate evidence using ViT [35] and BERT [45] and stores it in memory M. It takes a query q as input and retrieves its top-K nearest neighbors from the memory M containing image-text pairs. The retrieval results are then combined with the query q as enhanced input to the main encoder-decoder for answer generation. SKURG [8] combines evidence features using entity relations as carriers and inputs them into a transformer to generate key evidence and answers. This approach employs Bart [46] and OFA [36] as text and image feature encoders, respectively.

4.3. Implementation Details

In the evidence coarse screening stage, both our evidence and question encoding models are based on Bart-base [46]. During training, we utilize AdamW [47] as the optimizer and employ a linear strategy for learning rate decay. Additionally, in this stage, we set the top-K to 16. To address the scarcity of questions in the dataset, we augment the training set with data from SQuAD [3] and SQuAD2.0 [3]. For the image-to-text transformation process, we use OFA-large [36] to extract image descriptions and Fast-RCNN [37] to extract object information from the images. In the evidence fine screening stage, we employ Deberta-large [48] as the backbone. For training, we once again use AdamW as the optimizer. In the questionanswering component, we use mPlug-owl as the backbone, and during training, we maintain the use of AdamW as the optimizer. For a comprehensive overview of the experimental configurations, please refer to Table 2. In the table, the first column outlines the experimental settings for the Evidence Initial Screening Module (EISM), the second column for the Iterative Evidence Retrieval Module (IER), and the third column for the Question-Answering Module.

Additionally, it's worth noting that in the ablation experiments, after removing the EISM module, we randomly selected 16 candidates as the initial screening results.

Table 1: Overall Statistics of the downstream dataset.

Detecat	Train	Dev	Test		
Dataset	Image/Text/Table	Image/Text/Table	Image/Text/Table		
WebQA	18K/17K/0	2.5K/2.4K/0	3.4K/4K/0		
MultimodalQA	4.5K/12.6K/7.1K	0.8K/1.6K/0.9K	-		

Table 2: Hyperparameters of the different modules that were used on our method. The EISM means "Evidence Initial Screening Module". IER means the "Iterative Evidence Retrieval Module". Q&A means the "Question and Answering Module"

	EISM	IER	Q&A
Batch Size	256	8	32
Learning Rate	2×10^{-4}	2×10^{-5}	10^{-6}
Optimizer	AdamW	AdamW	AdamW
Scheduler	Linear	Linear	Linear

4.4. Main Results

Our results on WebQA [6] are presented in Table 3. "(Qonly)" refers to inference using only the questions. It can be observed that our approach outperforms the previous state-ofthe-art (SOTA) methods in all metrics. We surpass SOTA by 1.4% in the retrieval metric Retr-F1 and by 6.7% in the crucial QA metric. Additionally, we outperform SOTA by 6.3% in the QA fluency metric QA-FL and by 6.8% in the keyword accuracy metric. This demonstrates the robustness of our approach. Furthermore, it is noteworthy that our approach exhibits a substantial lead in the QA metrics, indicating that using initial evidence information rather than pooled features is more conducive to the accuracy of question answering. This also underscores the effectiveness of multi-turn QA systems in this task. In terms of retrieval, our F1 score reaches 89.6%, which is close to the human-evaluated F1 score of 90.5%. This further validates the effectiveness of our step-by-step evidence retrieval strategy for this task.

Our results on MultiModalQA [7] are presented in Table 4. "Multi-modal" refers to results for questions requiring multi-modal joint inference, "Single-modal" refers to results for questions requiring single-modal inference, and "ALL" represents results across the entire dataset. It can be observed that our approach outperforms the previous state-of-the-art (SOTA) methods in all metrics, showcasing the stability of our approach. Furthermore, our approach exhibits significant improvements over SOTA not only in multi-modal inference (EM: 2.2%, F1: 2.9%) but also in single-modal inference (EM: 3.6%, F1: 4.4%). This demonstrates that our method can enhance the performance of multi-modal retrieval question answering while maintaining high performance in single-modal retrieval question answering.

To perform a fair comparison with MuRAG [9] on Mul-

tiModalQA, we followed MuRAG's approach by selecting questions of two types, ImageQ and TextQ, from this dataset for testing. The results are shown in Table 5. It is evident that our method significantly outperforms the SOTA models on both of these question types. In comparison to the results on the complete dataset, our performance is even better on ImageQ and TextQ. This is because our model has not been specifically pretrained on TableQ, and its performance on this question type may not be as strong as on ImageQ and TextQ. Moreover, it may not perform as well on the challenging task of multi-modal inference as it does on single-modal question types.

4.5. Ablation Study

The results from the previous section indicate that PERQA outperforms its counterparts in challenging multi-modal retrieval tasks. In this section, we will further enhance understanding of our method through additional experiments and visualizations.

Effect of Negative Samples Semi-Supervised Contrastive Learning. We first investigated the impact of NSCL training on the initial screening model. On the validation set of WebQA, we used Recall@k to assess the model's ability to identify key sources from candidate sources. This metric represents the number of key sources identified among the top-k sources with the highest predicted scores divided by the total number of key sources. Finally, we selected the top two sources with the highest predicted scores and a score difference of less than 0.1 to evaluate the overall retrieval ability of this part on the WebQA test set. The experimental results, as shown in Table 7, indicate the following: The first row represents the performance of the model trained solely with the naive contrastive learning method for initial screening. The second row represents the performance after removing the negative sample semi-supervision and instead adding new data to train the model. The third row shows the performance achieved using only negative sample semisupervised contrastive learning. The fourth row represents the performance of the model after full training, including the utilization of the NSCL strategy and additional data during the training process. By comparing all the data, we can infer that increasing training data and utilizing the negative sample semi-supervised contrastive learning strategy both enhance the model's retrieval capabilities. Furthermore, by comparing the second and third rows, we can see that our proposed neg-

Table 3: Experiment results (%) on WebQA official test-set. The best results are in bold.

Model	QA-FL↑	QA-ACC↑	QA ↑	Retr-F1 ↑
VLP(Q-only) [2022]	34.9	22.2	13.4	-
VLP [2022]	42.6	36.7	22.6	68.9
VLP+VinVL [2022]	44.2	38.9	24.1	70.9
MuRAG [2022]	55.7	54.6	36.1	74.6
SKURG [2023]	55.4	57.1	37.7	88.2
PERQA (ours)	61.7	63.9	44.4	89.6

Table 4: Experiment results (%) on MultimodalQA dev-set. The best results are in bold.

Model	Multi-	modal	Single-	modal	ALL		
Model	EM ↑	F1 ↑	EM ↑	F1 ↑	EM ↑	F1 ↑	
AutoRoute [2020]	34.2	40.2	51.7	58.5	44.7	51.1	
ImplicitDecomp [2020]	44.6	51.2	51.6	58.4	48.8	55.5	
SKURG [2023]	52.5	57.2	66.1	69.7	59.8	64.0	
PERQA (ours)	54.7	60.3	69.7	74.1	62.8	67.8	

ative sample semi-supervised contrastive learning strategy is more effective than simply increasing training samples. However, we also observe that using only initial screening for retrieval does not perform well on the test set. This is because this method is challenging in terms of accurately determining the number of key sources and distinguishing them from indistinguishable distractor sources, making it difficult to improve precision.

Effect of Model Components on Overall Performance. We conducted a series of experiments on the MultimodalQA dataset to investigate the impact of different model components on overall performance. We used EM and F1 scores provided by the dataset as evaluation metrics for the questionanswering task and Retr-Pre, Retr-Rec, and Retr-F1 for evidence retrieval. Firstly, we compared our model with the existing state-of-the-art (SOTA) model SKURG [8] in MultiModalQA. The results showed that our approach outperformed the SOTA model in both the question-answering and retrieval tasks. It's worth noting that our model achieved slightly lower Retr-Pre scores than SKURG in the multimodal evidence retrieval task but surpassed SKURG in terms of Retr-Rec and Retr-F1 scores. Our analysis suggests that this is because our model sets lower standards when selecting key sources, allowing it to extract more key sources and thus improving recall. However, this approach also introduces more noise sources, leading to lower precision. Nevertheless, we believe that the model's performance in recall is more critical, as question-answering models can be trained to mitigate interference from noise sources, but without key sources, the model cannot correctly answer questions.

The third row (w/o NSCL) signifies that we retained the initial screening module but did not train the model with NSCL. When compared to the second row, the model's performance in both single-modal and multi-modal retrieval

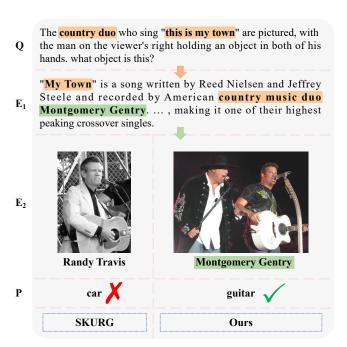


Fig. 5: Examples of retrieval Q&A. Q represents the question, E_1 represents the first evidence, E_2 represents the second evidence, and P represents the model prediction result.

and question-answering accuracy decreased, indicating that NSCL training is essential for this framework. The data in the fourth and fifth rows represent the model's accuracy after removing the evidence initial screening module and the iterative retrieval module, respectively. We can observe a significant drop in recall accuracy after removing EISM, demonstrating that the initial screening module not only effectively reduces the retrieval space for iterative evidence retrieval but also enhances overall retrieval correctness. By comparing the data

Table 5: Experiment results (%) on subsets of MultimodalQA dev-set for different question types. The best results are in bold.

Model	Te	xt	Ima	ige	Text-Image
Model	EM ↑	F1 ↑	EM ↑	F1 ↑	$EM \uparrow$
Question-only [2020]	15.4	18.4	11.0	15.6	13.8
AutoRoute [2020]	49.5	56.9	37.8	37.8	46.6
MuRAG [2022]	60.8	67.5	58.2	58.2	60.2
SKURG [2023]	66.7	72.7	56.1	56.1	64.2
PERQA (ours)	74.6	80.5	63.9	64.1	72.0

Table 6: Ablation Experiments Results (%) on the MultiModalQA Dataset. Here, EISM denotes the Evidence Initial Screening Module, NSCL signifies training the initial screening model using Negative Samples semi-supervised Contrastive Learning, and IER represents the Iterative Evidence Retrieval Module. The best results are in bold.

Row Model		Multi-modal						Single-modal				All				
Row	Model	EM	F1	Retr-Pre	Retr-Rec	Retr-F1	EM	F1	Retr-Pre	Retr-Rec	Retr-F1	EM	F1	Retr-Pre	Retr-Rec	Retr-F1
1	SKURG	52.5	57.2	86.1	75.7	80.6	66.1	69.7	94.7	80.2	86.7	59.8	64.0	89.6	77.7	83.2
2	PERQA (ours)	54.7	60.3	81.7	80.7	81.2	69.7	74.1	95.0	82.7	88.4	62.8	67.8	87.1	81.6	84.2
3	- w/o NSCL	51.6	57.6	76.0	68.2	71.9	68.1	72.8	91.8	81.1	86.1	60.5	65.8	82.8	73.8	78.0
4	- w/o EISM	45.0	50.7	68.8	60.6	64.5	54.9	57.9	61.2	53.0	56.8	50.4	54.6	65.5	57.3	61.2
5	- w/o IER	33.5	38.5	62.8	47.7	54.2	61.6	66.1	62.7	69.7	66.0	48.7	53.4	62.7	57.3	59.9
6	- w/o (NSCL + IER)	30.4	34.8	43.3	47.3	45.2	58.4	63.5	42.4	68.0	52.3	45.6	50.3	42.9	56.3	48.7
7	- w/o (EISM + IER)	17.7	20.6	10.7	9.5	10.1	24.7	28.4	6.1	8.3	7.0	21.5	24.8	8.2	9.0	8.6

Table 7: Ablation experiments results (%) on negative sample supervised contrastive learning (NSCL) method. "AUG" denotes the use of additional training data, while "NEG" signifies the utilization of negative samples for semi-supervised training. The best results are in bold.

Model	We	WebQA test set		
Model	Recall@8	Recall@12	Recall@16	Retr F1
NSCL w/o (AUG+NEG)	80.3	93.4	96.4	62.7
NSCL w/o NEG	91.2	95.8	97.9	64.5
NSCL w/o AUG	95.1	97.5	98.7	73.0
NSCL	96.7	98.7	99.4	74.0

in the fourth and fifth rows, we can also see that, compared to removing the EISM module, removing the IER module results in worse overall performance in multi-modal scenarios (EM decreased by 11.5%, Retr-F1 decreased by 10.3%), but better performance in single-modal scenarios (EM improved by 6.7%, Retr-F1 improved by 9.0%). This indicates that the IER module excels in multi-modal multi-hop retrieval, while the EISM module performs better in single-modal retrieval. Combining both modules allows for effective complementarity. The sixth row represents our removal of the IER module, retaining only the initial screening module, and training the model with regular contrastive learning, while the seventh row represents the complete removal of all screening stages and random selection of key evidence from candidate sources for question-answering. Compared to the second row, there is a significant decline in model performance, highlighting the outstanding performance of our proposed method. The question-answering performance of the model that randomly selects key evidence from candidate sources is also less than ideal, emphasizing the necessity of the retrieval

stage. Through this series of ablation experiments, we further demonstrate the effectiveness of our proposed method.

4.6. Case Study

To demonstrate that our model can complete retrieval and answer inference by mining logical order relationships among pieces of evidence, we have provided an illustrative example. As shown in Figure 5, we present the retrieval and inference process of our method (on the right) and the state-of-the-art (SOTA) SKURG [8] method (on the left) for a specific question. In this figure, E_1 represents intermediate key evidence, which is associated with the question through the orange keywords and linked to other key evidence through the green keywords. E_2 is the final key evidence, crucial for reasoning towards the answer. The model needs to correctly retrieve E_2 to answer the question. We can observe that both our method and SKURG are able to retrieve the intermediate key evidence E_1 . However, SKURG incorrectly estimates the final key evidence E_2 , while our method successfully retrieves it, leading to a correct answer in our case and an incorrect one in SKURG's case. This example illustrates how our method leverages potential logical order relationships among key evidence to retrieve them and answer questions accurately.

5. CONCLUSION

In this work, we propose a two-stage evidence retrieval question-answering method. Our proposed method, named PERQA, utilizes a progressive evidence refinement strategy for evidence retrieval and employs an iterative evidence retrieval approach to elucidate the logical sequence among pieces of evidence. Additionally, we introduce a semisupervised training strategy based on negative samples to fully exploit latent information in known samples within our retrieval model. Finally, we design a multi-turn questionanswering strategy to handle multi-modal inputs, incorporating multiple key pieces of evidence as historical context along with the question input into the model and leveraging cross-modal attention mechanisms to capture potential connections between evidence and questions. Through extensive experiments, we demonstrate the outstanding performance of our method in multi-modal retrieval question-answering tasks, and we look forward to providing valuable insights for future research.

6. REFERENCES

- [1] N Donratanapat, S Samadi, José M Vidal, and S Sadeghi Tabas, "A national scale big data analytics pipeline to assess the potential impacts of flooding on critical infrastructures and communities," *Environmental Modelling & Software*, vol. 133, pp. 104828, 2020.
- [2] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh, "Vqa: Visual question answering," in *International Conference on Computer Vision (ICCV)*, 2015, pp. 2425–2433.
- [3] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang, "Squad: 100,000+ questions for machine comprehension of text," in *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2016, pp. 2383–2392.
- [4] Pranav Rajpurkar, Robin Jia, and Percy Liang, "Know what you dont know: Unanswerable questions for squad," in *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 2018, pp. 784–789.
- [5] Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi, "Aokvqa: A benchmark for visual question answering using world knowledge," in *European Conference on Computer Vision (ECCV)*. Springer, 2022, pp. 146–162.
- [6] Yingshan Chang, Mridu Narang, Hisami Suzuki, Guihong Cao, Jianfeng Gao, and Yonatan Bisk, "Webqa: Multihop and multimodal qa," in *Conference on Com*puter Vision and Pattern Recognition (CVPR), 2022, pp. 16495–16504.
- [7] Alon Talmor, Ori Yoran, Amnon Catav, Dan Lahav, Yizhong Wang, Akari Asai, Gabriel Ilharco, Hannaneh Hajishirzi, and Jonathan Berant, "Multimodalqa: complex question answering over text, tables and images," in

- International Conference on Learning Representations (ICLR), 2020.
- [8] Qian Yang, Qian Chen, Wen Wang, Baotian Hu, and Min Zhang, "Enhancing multi-modal and multi-hop question answering via structured knowledge and unified retrieval-generation," ACM International Conference on Multimedia (ACMMM), 2023.
- [9] Wenhu Chen, Hexiang Hu, Xi Chen, Pat Verga, and William Cohen, "Murag: Multimodal retrievalaugmented generator for open question answering over images and text," in *Conference on Empirical Methods* in *Natural Language Processing (EMNLP)*, 2022, pp. 5558–5570.
- [10] Feng Gao, Qing Ping, Govind Thattai, Aishwarya Reganti, Ying Nian Wu, and Prem Natarajan, "Transform-retrieve-generate: Natural language-centric outside-knowledge visual question answering," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022, pp. 5067–5077.
- [11] Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih, "Dense passage retrieval for open-domain question answering," in *Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 2020.
- [12] Michael Glass, Gaetano Rossiello, Md Faisal Mahbub Chowdhury, Ankita Rajaram Naik, Pengshan Cai, and Alfio Gliozzo, "Re2g: Retrieve, rerank, generate," in Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL), 2022.
- [13] Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton, "A simple framework for contrastive learning of visual representations," in *Proceedings of International Conference on Machine Learning (ICML)*. PMLR, 2020, pp. 1597–1607.
- [14] Tianyu Gao, Xingcheng Yao, and Danqi Chen, "Simcse: Simple contrastive learning of sentence embeddings," in *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2021, pp. 6894–6910.
- [15] Chenliang Li, Haiyang Xu, Junfeng Tian, Wei Wang, Ming Yan, Bin Bi, Jiabo Ye, He Chen, Guohai Xu, Zheng Cao, et al., "mplug: Effective and efficient vision-language learning by cross-modal skipconnections," in *Conference on Empirical Methods* in Natural Language Processing (EMNLP), 2022, pp. 7241–7259.

- [16] Qinghao Ye, Haiyang Xu, Guohai Xu, Jiabo Ye, Ming Yan, Yiyang Zhou, Junyang Wang, Anwen Hu, Pengcheng Shi, Yaya Shi, et al., "mplug-owl: Modularization empowers large language models with multimodality," *arXiv preprint arXiv:2304.14178*, 2023.
- [17] E VOORHEES, "The trec-8 question answering track report," in *Proceedings of the Text Retrieval Conference* (*TREC*), 1999.
- [18] Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes, "Reading wikipedia to answer open-domain questions," in *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2017, pp. 1870–1879.
- [19] Wei Yang, Yuqing Xie, Aileen Lin, Xingyu Li, Luchen Tan, Kun Xiong, Ming Li, and Jimmy Lin, "End-to-end open-domain question answering with bertserini," in *Annual Conference of the North American Chapter of the Association for Computational Linguistics* (NAACL), 2019, pp. 72–77.
- [20] Yixin Nie, Songhe Wang, and Mohit Bansal, "Revealing the importance of semantic retrieval for machine reading at scale," in *Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 2019, pp. 2553–2566.
- [21] Tomer Wolfson, Mor Geva, Ankit Gupta, Matt Gardner, Yoav Goldberg, Daniel Deutch, and Jonathan Berant, "Break it down: A question understanding benchmark," *Transactions of the Association for Computational Linguistics (TACL)*, vol. 8, pp. 183–198, 2020.
- [22] Sewon Min, Danqi Chen, Luke Zettlemoyer, and Hannaneh Hajishirzi, "Knowledge guided text retrieval and reading for open domain question answering," *arXiv* preprint arXiv:1911.03868, 2019.
- [23] Akari Asai, Kazuma Hashimoto, Hannaneh Hajishirzi, Richard Socher, and Caiming Xiong, "Learning to retrieve reasoning paths over wikipedia graph for question answering," in *International Conference on Learning Representations (ICLR)*, 2019.
- [24] Sean MacAvaney, Andrew Yates, Arman Cohan, and Nazli Goharian, "Cedr: Contextualized embeddings for document ranking," in *Annual ACM Conference on Research and Development in Information Retrieval*, 2019, pp. 1101–1104.
- [25] Dongmei Chen, Sheng Zhang, Xin Zhang, and Kaijing Yang, "Cross-lingual passage re-ranking with alignment augmented multilingual bert," *IEEE Access*, vol. 8, pp. 213232–213243, 2020.

- [26] Kai Zhang, Chongyang Tao, Tao Shen, Can Xu, Xiubo Geng, Binxing Jiao, and Daxin Jiang, "Led: Lexiconenlightened dense retriever for large-scale retrieval," in *Proceedings of the ACM Web Conference (ACMWeb)*, 2023, pp. 3203–3213.
- [27] Cen Chen, Chengyu Wang, Minghui Qiu, Dehong Gao, Linbo Jin, and Wang Li, "Cross-domain knowledge distillation for retrieval-based question answering systems," in *Proceedings of the ACM Web Conference* (ACMWeb), 2021, pp. 2613–2623.
- [28] Wei-Cheng Chang, X Yu Felix, Yin-Wen Chang, Yiming Yang, and Sanjiv Kumar, "Pre-training tasks for embedding-based large-scale retrieval," in *International Conference on Learning Representations (ICLR)*, 2019.
- [29] Omar Khattab and Matei Zaharia, "Colbert: Efficient and effective passage search via contextualized late interaction over bert," in *Annual ACM Conference on Research and Development in Information Retrieval*, 2020, pp. 39–48.
- [30] Yingqi Qu, Yuchen Ding, Jing Liu, Kai Liu, Ruiyang Ren, Wayne Xin Zhao, Daxiang Dong, Hua Wu, and Haifeng Wang, "Rocketqa: An optimized training approach to dense passage retrieval for open-domain question answering," in *Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, 2021, pp. 5835–5847.
- [31] Ruiyang Ren, Yingqi Qu, Jing Liu, Wayne Xin Zhao, Qiaoqiao She, Hua Wu, Haifeng Wang, and Ji-Rong Wen, "Rocketqav2: A joint training method for dense passage retrieval and passage re-ranking," in Conference on Empirical Methods in Natural Language Processing (EMNLP), 2021, pp. 2825–2835.
- [32] Keshav Santhanam, Omar Khattab, Jon Saad-Falcon, Christopher Potts, and Matei Zaharia, "Colbertv2: Effective and efficient retrieval via lightweight late interaction," in Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL), 2022, pp. 3715–3734.
- [33] Darryl Hannan, Akshay Jain, and Mohit Bansal, "Manymodalqa: Modality disambiguation and qa over diverse inputs," in *The AAAI Conference on Artificial Intelligence (AAAI)*, 2020, vol. 34, pp. 7879–7886.
- [34] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 25, 2012.
- [35] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas

- Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al., "An image is worth 16x16 words: Transformers for image recognition at scale," in *International Conference on Learning Representations (ICLR)*, 2020.
- [36] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang, "Ofa: Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework," in *Proceedings of International Conference on Machine Learning (ICML)*. PMLR, 2022, pp. 23318–23340.
- [37] Ross Girshick, "Fast r-cnn," in *International Conference on Computer Vision (ICCV)*, 2015, pp. 1440–1448.
- [38] Rodrigo Nogueira and Kyunghyun Cho, "Passage reranking with bert," *arXiv preprint arXiv:1901.04085*, 2019.
- [39] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al., "Llama: Open and efficient foundation language models," arXiv preprint arXiv:2302.13971, 2023.
- [40] Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al., "Lora: Low-rank adaptation of large language models," in *International Conference on Learning Representa*tions (ICLR), 2021.
- [41] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al., "Learning transferable visual models from natural language supervision," in *Proceedings of International Conference on Machine Learning (ICML)*. PMLR, 2021, pp. 8748–8763.
- [42] Weizhe Yuan, Graham Neubig, and Pengfei Liu, "Bartscore: Evaluating generated text as text generation," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 34, pp. 27263–27277, 2021.
- [43] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv* preprint arXiv:1907.11692, 2019.
- [44] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee, "Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks," *Advances in Neural Information Processing Systems (NeurIPS)*, vol. 32, 2019.

- [45] Jacob Devlin Ming-Wei Chang Kenton and Lee Kristina Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL), 2019, pp. 4171–4186.
- [46] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer, "Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension," in *Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020, pp. 7871–7880.
- [47] Ilya Loshchilov and Frank Hutter, "Decoupled weight decay regularization," in *International Conference on Learning Representations (ICLR)*, 2018.
- [48] Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen, "Deberta: Decoding-enhanced bert with disentangled attention," in *International Conference on Learning Representations (ICLR)*, 2020.