electronics

MDPI

Article

Prompt-Enhanced Generation for Multimodal Open Question Answering

Chenhao Cui [1] and Zhoujun Li [2,*]

[1] School of Cyber Science and Technology, Beihang University, Beijing 100191, China; cuich@buaa.edu.cn
[2] School of Computer Science and Engineering, Beihang University, Beijing 100191, China
* Correspondence: lizj@buaa.edu.cn

Abstract: Multimodal open question answering involves retrieving relevant information from both images and their corresponding texts given a question and then generating the answer. The quality of the generated answer heavily depends on the quality of the retrieved image–text pairs. Existing methods encode and retrieve images and texts, inputting the retrieved results into a language model to generate answers. These methods overlook the semantic alignment of image–text pairs within the information source, which affects the encoding and retrieval performance. Furthermore, these methods are highly dependent on retrieval performance, and poor retrieval quality can lead to poor generation performance. To address these issues, we propose a prompt-enhanced generation model, PEG, which includes generating supplementary descriptions for images to provide ample material for image–text alignment while also utilizing vision–language joint encoding to improve encoding effects and thereby enhance retrieval performance. Contrastive learning is used to enhance the model's ability to discriminate between relevant and irrelevant information sources. Moreover, we further explore the knowledge within pre-trained model parameters through prefix-tuning to generate background knowledge relevant to the questions, offering additional input for answer generation and reducing the model's dependency on retrieval performance. Experiments conducted on the WebQA and MultimodalQA datasets demonstrate that our model outperforms other baseline models in retrieval and generation performance.

Keywords: multimodal question answering; retrieval augmented generation; prompt learning; vision–language alignment

Citation: Cui, C.; Li, Z. Prompt-Enhanced Generation for Multimodal Open Question Answering. *Electronics* **2024**, *13*, 1434. https://doi.org/10.3390/electronics13081434

Academic Editor: Arkaitz Zubiaga

Received: 14 March 2024
Revised: 7 April 2024
Accepted: 8 April 2024
Published: 10 April 2024

1. Introduction

Multimodal question answering represents a significant advancement in the field of artificial intelligence, merging computer vision, natural language processing, and knowledge representation. This integration allows models to understand and process information from both visual and textual inputs, mimicking human-like abilities to interpret and respond to complex queries that require cross-modal comprehension and reasoning. Multimodal question answering datasets [1,2] require models to answer questions based on given images and texts. Furthermore, multimodal open question answering datasets [3,4] require models to retrieve relevant evidence from image and text information sources based on the question and integrate evidence to generate answers.

Existing works follow a retrieval and generation pipeline. Some works [3,5] concatenate the question with information sources and evaluate the probability of information sources being selected. The question and the selected information sources are then concatenated and input into the model to generate an answer to the question. Methods [2,4] initially employ a question-type classifier to identify the modality (image or text) that may contain the answer. Subsequently, the question and input sources are allocated to different unimodal sub-models to generate answers. MuRAG [6] uses a joint encoder for images and

text and retrieves the top-k nearest neighbors from the memory of image–text pairs based on the question.

Previous methods, when encoding images and texts, overlook the semantic alignment between images and text. As shown in Figure 1, the description of the image is usually information such as the location and time of the image, which the image itself cannot describe and is not related to the content of the image. Whether encoding separately or jointly across modalities, the inability to semantically align images and their descriptive texts affects the semantic interaction alignment between them. This results in the image–text pairs not being well associated with the questions, thereby reducing encoding and retrieval performance. Furthermore, these models use the retrieved results as the only question-related information input, resulting in a heavy dependence on the quality of retrieval. If irrelevant images and texts are retrieved, the accuracy of the model-generated answers will significantly decrease. Recent research has shown that large pre-trained model parameters carry world knowledge [7,8], which can be applied to downstream tasks through appropriate prompt learning [9,10]. Mining question-related background knowledge from the information in pre-trained model parameters can mitigate the issue of answer generation is heavily dependent on the quality of retrieved information sources.

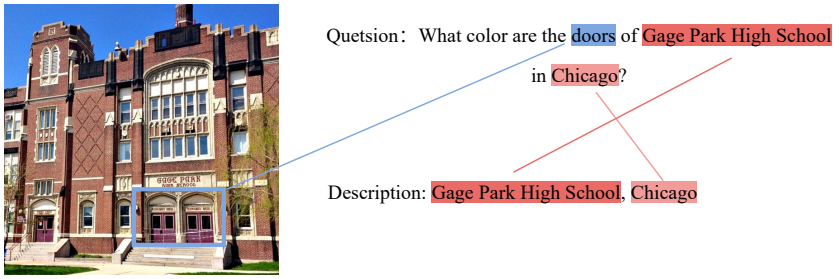

Figure 1. Example of image description in webQA dataset. While the image and description correspond to elements in the question, there is no semantic relevance between them.

To address the limitations mentioned above, we propose a prompt-enhanced generation model, PEG, for multimodal open question answering. First, to address the issue of insufficient image descriptions that do not easily semantically correspond to the image content, we employ prompt learning to generate supplementary descriptions for images, ensuring semantic consistency between image and text content. Subsequently, we use vision–language joint encoding to finely align the semantics of images and texts, enhancing the representations of both vision and language. Improvement in encoding capacity, in turn, leads to improved retrieval results. To enhance the model's ability to discriminate information sources, we treat question-related information sources as positive examples and irrelevant ones as negative examples, adopting contrastive learning to further enhance the model's representation and discrimination capabilities. To address the issue of heavy reliance on retrieval quality, we utilize prefix-tuning in prompt learning to mine background knowledge contained within model parameters, providing additional information input for answer generation. This enables the model to generate answers based on background knowledge, even when irrelevant information is retrieved.

Our contributions can be summarized as follows:

- We propose a prompt-enhanced generation model, PEG, for multimodal open question answering. By employing prompt learning to generate supplementary descriptions for images, ample material is provided for semantic alignment of vision and language. Subsequently, vision–language joint encoding is utilized for fine-grained alignment of visual and linguistic semantics, thereby enhancing encoding and retrieval performance.
- We employ prefix-tuning to mine the knowledge within pre-trained model parameters, generating background knowledge related to the question. This provides an additional input for answer generation, thereby reducing the model's dependency on retrieval

results. Thus, the model can generate answers to the question, even if it retrieves irrelevant information.
- The experimental results for two datasets demonstrate that our model, utilizing prompt learning, can enhance encoding and retrieval capabilities, generating more accurate answers.

2. Related work

2.1. Multimodal Question Answering

Antol et al. [11] first introduced the Visual Question Answering (VQA) task, which requires models to answer questions based on image inputs. OK-VQA [12] extends the task to knowledge-seeking questions, where the content of the image alone is insufficient to answer the question, encouraging models to rely on external knowledge resources. In ManyModalQA [1], the context for each question includes information from multiple modalities. However, the answer to each question can only be derived from a single modality, eliminating the need for cross-modal reasoning. MuMuQA [2] requires models to perform reasoning within given images and text without the need for retrieval. MIMOQA [13] proposes a new concept of multimodal input and multimodal output, which requires the model to output text answers and a related image. The aforementioned datasets do not require models to perform retrieval. They only need to process the given text and images, resembling multimodal input reading comprehension rather than open-book question answering. WebQA [3] and MultimodalQA [4] are two multimodal open-book question answering datasets. Models are required to retrieve relevant images and texts from information sources based on the question, and through logical or numerical reasoning, aggregate information from multiple sources to generate an answer to the question.

Some works [3,5] concatenate the question with information sources and evaluate the probability of information sources being selected. The question and the selected information sources are then concatenated and input into the model to generate an answer to the question. Methods [2,4] initially employ a question-type classifier to identify the modality (image or text) that may contain the answer. Subsequently, the question and input sources are allocated to different unimodal sub-models to generate answers. MuRAG [6] uses the maximum inner product to retrieve the top-k nearest neighbors from the memory of image–text pairs based on the question. The retrieved results and the question are concatenated and input into an encoder–decoder to generate an answer. Lin et al. [14] convert the image into text and propose a joint training scheme which includes differentiable DPR integrated with answer generation so that the system can be trained in an end-to-end fashion. Yang et. al. [15] convert the image into captions (or tags) that GPT-3 can understand then adapt GPT-3 to solve the VQA task in a few-shot manner by just providing a few in-context VQA examples. Yu et al. [16] convert images and tables into a unified linguistic representation, thereby simplifying the task into a simpler text-based QA problem, which is solved using three steps: retrieval, ranking, and generation. Research [17] has been conducted to explore using prompts with GPT-4 to solve multimodal question-answering.

2.2. Retrieval Augmented Generation

Retrieval-Augmented Generation combines parametric models with external relevant knowledge to inject world knowledge into language models, thereby generating more accurate and knowledge-rich answers. Dense Passage Retrieval (DPR) [18] includes a question encoder and a document encoder for encoding questions and documents into separate dense representations. The training objective of the DPR is to assign higher scores to documents that can help answer the question, thereby enabling the retrieval of a set of documents to be passed to the answer generation module. REALM [19] is an encoder-only model that integrates Wikipedia as a database to support downstream knowledge-intensive tasks. RAG [20] utilizes a pre-trained BART model as its reader, generating answers by inputting the question along with documents retrieved by DPR. FID [21] first independently encodes each retrieved document using a BART encoder, then uses a decoder to perform attention calculation on all output representations to generate the final answer. We will

apply retrieval enhanced generation to multimodal open question answering, using vision and language joint encoding to retrieve knowledge from both image and text modalities, rather than only text knowledge.

3. Method

3.1. Task Formulation

Given a question Q and input information sources $\mathbf{S} = \{S_1, S_2, \ldots, S_n\}$, where n represents the number of information sources, each information source can be a text with a title or an image with captions. The task's target is to retrieve relevant source data from $\mathbf{S}$ based on the question Q and generate a natural language answer $A = a_1, a_2, \ldots, a_n$, where a_i is a word in the answer sentence. $\hat{S}$ is the result retrieved based on question Q from $\mathbf{S}$, represented as:

$$\hat{S} = Top_K(\mathbf{S} \mid Q) \tag{1}$$

The task objective is to train a neural network generative model $p(A \mid Q, \hat{S})$ that can generate an answer $\hat{A}$ that is both consistent with the retrieved source data $\hat{S}$ and fluent.

3.2. Model Overview

We show the main architecture of our PEG model in Figure 2. For the input information sources of a given question, an image usually includes a text description. However, the text description often serves as a supplement to the image, and their semantic information does not correspond directly. Therefore, we first generate a supplementary description for the image through prompt-based learning, providing a corpus for full alignment of visual and linguistic information. Since the input information sources include image–text pairs, we adopt a joint vision–language encoding approach to capture the features of both images and texts. The information source index refers to the encoded vector database of the information source (image–text pair), which facilitates the use of question vectors to retrieval related information sources through MIPS. To further provide the background knowledge required by the question, we utilize the knowledge contained in the parameters of large pre-trained models and generate background knowledge related to the question using prefix-tuning as an additional input relevant to the question. After obtaining the representations of the question and information sources, we retrieve information related to the question from the information sources. The question, relevant information sources, and background knowledge are then fed into an answer generator to produce the answer.

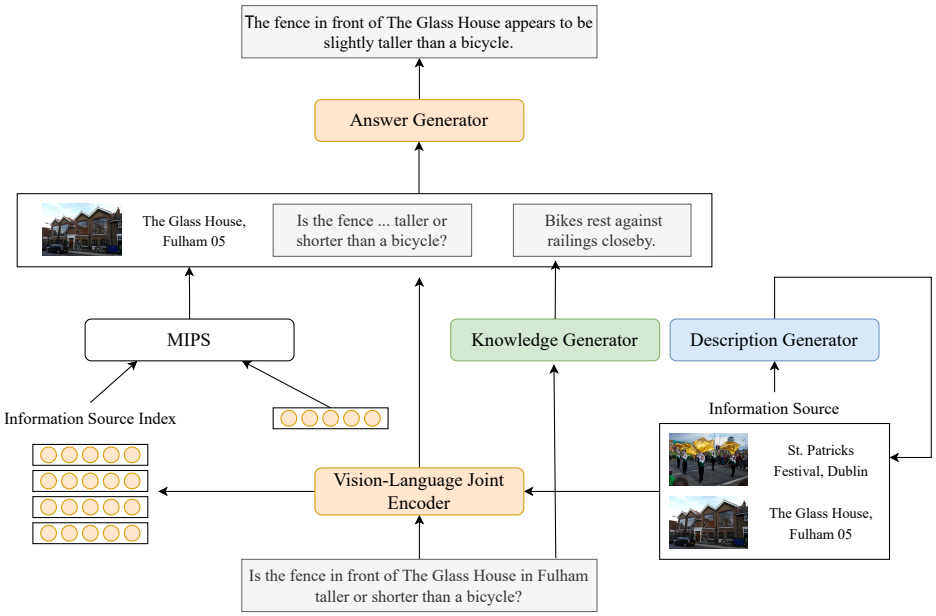

Figure 2. The main architecture of our proposed PEG model.

3.3. Prompt-Based Image Supplementary Description Generation

For each image–text pair in the input information sources, the text is often a very simple description of the image, typically the location and time associated with the image, as shown in Figure 3. Due to the lack of semantic connection between the text descriptions and the image content, there can be insufficient alignment between the semantics of images and texts. This misalignment can lead to poor encoding effectiveness for image–text pairs, which in turn affects the retrieval performance. To better describe the content of the images, we first use the pre-trained caption generation model OFA [22] to generate descriptions of the images.

Figure 3. Example of image supplementary description in webQA dataset.

OFA is a large, multimodal model that unifies language tasks, visual tasks, and tasks involving both vision and language, encompassing both multimodal understanding and generation tasks such as text-to-image generation, visual question answering, and visual reasoning. The model employs a sequence-to-sequence learning architecture with task representation based on prompts. For visual inputs, OFA utilizes a ResNet module to convert images into feature representations through convolutional operations. For textual inputs, OFA employs Byte Pair Encoding (BPE) to transform given text sequences into subword sequences and then into embedding features. OFA uses a Transformer as its backbone architecture, with an encoder–decoder framework serving as the unified structure for all tasks.

OFA sets different prompts for different tasks. For the image captioning task, the model generates descriptions of the image based on the given image and the prompt "What does the image describe?". The generated supplementary descriptions are relatively consistent with the image content, such as elements like school, buildings, bricks, windows, doors, etc., facilitating the alignment of image and text semantics in the next step of vision–language joint representation.

3.4. Vision–Language Joint Representation

The vision–language joint encoder we used is shown in Figure 4. Firstly, we employ a backbone network as the image tokenizer to convert images into visual token embeddings in Figure 4b. Then, text embeddings and visual token embeddings are concatenated as the input of the main encoder–decoder framework in Figure 4a.

Text Embeddings: Each text in the information sources is firstly converted into the sequence of tokens $\{t_1, t_2, \ldots, t_T\}$ and then two special tokens "$\langle s \rangle$" and "$\langle \backslash s \rangle$" are added to represent the start and end of the document. After that, we map each token into vector representation $E_T = \{e_{start}, e_1, \ldots, e_T, e_{end}\}$ with the text-embedding layer.

Image Embeddings: Different from previous methods, which extract many image features via existing object detection models, we employ ViT [23] as the backbone, which splits each image into several patches and then encodes them. The details of the image tokenizer are shown in Figure 4b.

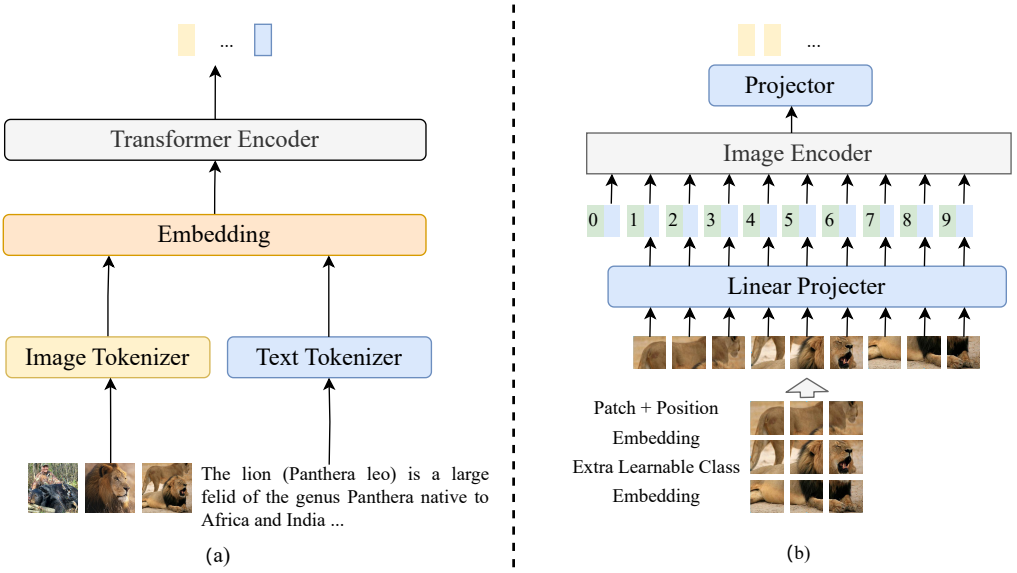

Figure 4. (**a**) is Vision–language joint encoder architecture and (**b**) represents the image tokenizer.

Firstly, we reshape image $img \in \mathbb{R}^{H\times W\times C}$ into a sequence of flattened 2D patches $\{img^p \in \mathbb{R}^{N\times(P^2\cdot C)}\}_{p=1}^{N}$, where (H, W) is the resolution of the original image, C is the number of channels, (P, P) is the resolution of each image patch, and $N = HW/P^2$ is the resulting number of patches. Then, we can obtain a sequence of image patches $\{img^p\}_{p=1}^{N}$ as the input of the image tokenizer.

Secondly, the patches are linearly projected to patch embeddings $e^p = E \times img_i^p$, where $E \in \mathbb{R}^{(P^2\cdot D)\times C}$. We also add a special token "[class]" with learnable embedding e^0. Then, position embeddings and patch embeddings are attached as input Z_0 for the image encoder to retain positional information of images:

$$Z_0 = [e_i^0; e_i^1; \ldots; e_i^N] + E_{pos} \tag{2}$$

where $Z_0, E_{pos} \in \mathbb{R}^{(N+1)\times D}$, and E_{pos} are position embeddings.

Finally, we employ the pre-trained ViT as the backbone to encode these patches of each image. This backbone also can be replaced by any other encoders (e.g., a linear projection layer).

$$Z_{\ell+1} = \texttt{TransformerEncoderLayer}(Z_\ell), \ell = 1, 2, ..., L \tag{3}$$

The global max-pooling of output vectors is obtained as the visual token embedding E_i of image img_i, where $E_i \in R^D$.

Multi-Modal Encoder: The input of the multi-modal encoder is the concatenation of visual token embeddings E_i and token embeddings E_T. We can formalize the input as $H_0 = \{E_i; E_D\}$ and then encode visual and text embeddings with 12 transformer blocks. Finally, we can obtain vision–language representation H_L from the last layer output of this encoder.

$$H_L = \{h_{v_1}, \ldots, h_{v_M}, h_{start}, h_1, \ldots, h_{end}\} \tag{4}$$

The vision and language semantics interact with the self-attention mechanism of the transformer structure during the encoding process. It should be noted that the encoder takes image–text pairs as input, but either the image or text sequence can be empty to accommodate scenarios where only text or only images are provided as input.

3.5. Information Source Retrieval

The retriever is implemented based on DPR [18]. Given a question Q and input information sources $\mathbf{S}$ encoded by the vision–language joint encoder to obtain their vector representations, Maximum Inner Product Search (MIPS) [24] is used to iterate through $S \in \mathbf{S}$ to find the top-k input sources $\hat{S}$:

$$\hat{S} = \{S_1, S_2, \ldots, S_k\} = \underset{S \in \mathbf{S}}{Top_K} Encoder(Q) \cdot Encoder(S) \tag{5}$$

In the information sources $\mathbf{S}$ for a given question Q, there are inputs S_p that are relevant to the question and distracting inputs S_n that are irrelevant. (The datasets include annotations indicating whether the information source is relevant to the question.) By treating the input source S_p as a positive example and S_n as a negative example, a contrastive learning loss L_{con} can be constructed as follows:

$$L_{con} = -log \frac{exp\big(Encoder(Q) \cdot Encoder\big(S_p\big)\big)}{\sum\limits_{S \in \mathbf{S}} exp(Encoder(Q) \cdot Encoder(S))} \tag{6}$$

In the answer generation phase, the question Q and the retrieved sources $\hat{S}$ are concatenated $\{S_1, S_2, \ldots, S_k, Q\}$ as input. They are then encoded by the vision–language joint encoder to obtain a retrieval-enhanced vector representation.

3.6. Prefix-Tuning for Background Knowledge Generation

Recent studies have shown that large pre-trained models (PrLMs) contain a wealth of information within their vast number of parameters. This information can be leveraged for downstream tasks through the use of appropriate prompts, achieving notable results [25]. To further enhance the quality of answer generation, we explore the use of prefix-tuning to mine the parameter information within PrLMs, generating background knowledge related to the question as supplementary input for answer generation. Prefix-tuning [26] is a lightweight training paradigm that freezes the model parameters of PrLMs during training and only trains task-specific prefixes, avoiding the need to train all parameters of the large pre-trained model.

The knowledge generator takes a question Q as input and outputs background knowledge K related to the question. During the training process, the text T from the information source $S_p = (I, T)$ related to the question is used as the target knowledge K. As shown in Figure 5, the knowledge generator selects a pre-trained model with an encoder–decoder architecture and adds prefixes $[Prefix_0; Q; Prefix_1; K]$ in both the encoder and decoder, where Pre_0 and Pre_1 are the prefixes for the encoder and decoder, respectively. The prefix at the encoder side can guide the encoding of the input sequence, while the prefix at the decoder side can guide the generation of the subsequent target sequence. To avoid training instability and performance degradation caused by direct updates to the prefix parameters Pre_θ, a mapping matrix E_θ and a feedforward neural network MLP_θ are added, represented as:

$$Pre_\theta = MLP_\theta(E_\theta(Prefix)) \tag{7}$$

The loss function for background knowledge generation is represented as follows:

$$L_K = -\sum_{j=1}^{|K|} \log P(k_j | k_{<j}; \theta_{LM_K}, \theta_K) \tag{8}$$

where θ_{LM_K} represents the parameters of the knowledge generation model, which remain unchanged during training, and θ_K represents the learnable prefix parameters.

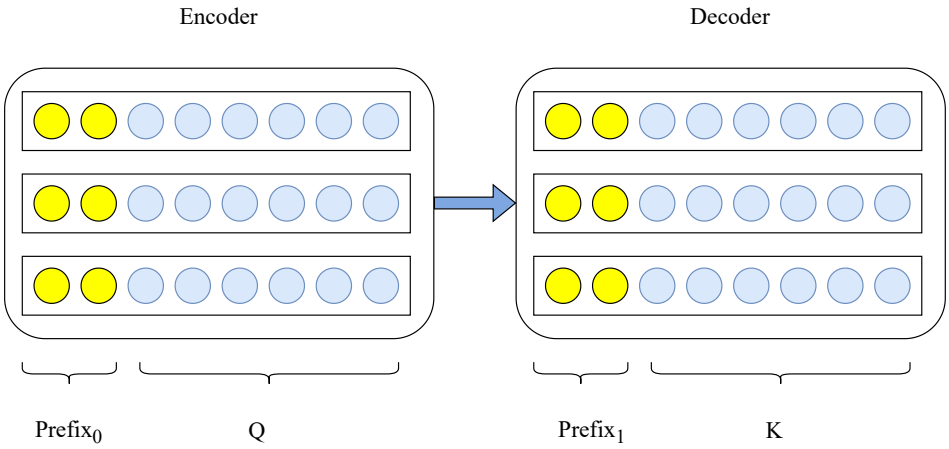

Figure 5. Encoder–decoder architecture prefix-tuning.

3.7. Answer Generation

We employ a generative model such as BART to generate answers and modify the encoder of BART to serve as the vision–language joint encoder. The input to the model comprises the concatenation of the question, retrieved information sources, and generated background knowledge. For a given question Q, through information source retrieval, we obtain related information sources $\hat{S} = \{S_1, S_2, \ldots, S_k\}$. By inputting the question into the knowledge generator, we obtain background knowledge K related to the question. We then concatenate these elements $\{Q; I_1, T_1; \ldots; I_k, T_k; K\}$ as the input. After encoding by the vision–language joint encoder, we pass the obtained representation H_L to the decoder to generate an answer $A = a_1, a_2, \ldots, a_n$, where a_j is a word in the answer sentence. The answer token is generated by the token with questions Q, related information sources $\hat{S}$, and background knowledge K as input, the loss function for the answer generation task is as follows:

$$L_A = -\sum_{j=1}^{|A|} \log P_\theta(a_j | a_{<j}; Q, \hat{S}, K) \tag{9}$$

where $|A|$ denotes the length of the answer, a_j represents the j-th token, P_θ represents the parameters of the model.

We use the answer generation task, contrastive learning for relevant information sources, and background knowledge generation task to jointly train our PEG model. In previous sections, we have introduced their details. Finally, PEG is trained with three tasks by simultaneously minimizing the following three loss functions:

$$L = L_A + L_{con} + L_K \tag{10}$$

4. Experiments

To verify the effectiveness of our model, this section will detail the experiments conducted and related specifics. It will be divided into subsections on relevant datasets, experimental design, baseline models, evaluation metrics, experimental results, and analysis for a comprehensive introduction.

4.1. Datasets

We utilized two representative datasets in the field of multimodal question answering: the WebQA [3] and MultimodalQA [4] datasets.

- WebQA is a multimodal, multi-hop reasoning dataset where all questions require the retrieval of one to two images or text segments as knowledge for answering. Each question in the dataset is associated with a set of visual and textual distractors, necessitating the retrieval of relevant visual and textual sources as knowledge input.

The answers in this dataset are typically complete sentences, facilitating a better evaluation of the model's generative capabilities.

- MultimodalQA contains multimodal question-answer pairs in text, image, and table formats. This dataset features 16 types of questions, most of which require multimodal retrieval and reasoning. Unlike WebQA, the answers in MultimodalQA are in the form of text fragments or phrases. The dataset initially generates questions through templates, then asks crowd workers to filter and interpret the generated questions. Each question is also associated with a set of visual and textual distractors. Following MuRAG [6], we have only considered text and image questions.

4.2. Implementation Details

Image description generation employed the OFA-large model (https://github.com/OFA-Sys/OFA (accessed on 9 April 2024)). The vision–language joint encoder and the decoder for answer generation were initialized using ViT-B/16 and BART-base [27], respectively, published on Hugging Face (https://huggingface.co/ (accessed on 9 April 2024)). The image tokenizer used the same configuration as in ViT. The batch size used during training was set to 16, the number of retrieved information sources (top-k) was set to 4, and the decoding process used beam search with a beam size of 3. The background knowledge generator was initialized with BART-large, with the prefix sequence length set to 20. The model training process was optimized using the Adam [28] algorithm and a Linear Learning Rate Scheduler. The model was trained on four NVIDIA Tesla V100 32G GPUs.

4.3. Baselines

We compare our approach with the following baseline models with WebQA and MultimodalQA datasets:

- Question-Only: This method does not search for information sources related to the question but directly inputs the question into the BART model to generate an answer.
- VLP [5]: Models such as Oscar [29] and VinVL [30] are typically used as standard baseline models. These models are based on transformer encoder–decoder architectures. They concatenate the question and information sources and evaluate the probability of information sources being selected. The question and the selected information sources are concatenated and inputted into the model to generate answers to the question.
- AutoRouting [4]: This method initially employs a question-type classifier to identify the modality (image or text) that may contain the answer. Subsequently, the question and input sources are allocated to different unimodal sub-models to generate answers, transforming multimodal question answering into several unimodal question answering tasks. This approach uses the RoBERTa [31] model to answer text questions, VilBERT [32] to answer image questions, and FasterRCNN [33] to extract image features.
- CLIP [34]: This model is used to encode questions and image–text input sources separately. It employs Approximated Nearest Neighbor Search (ANN) to find k candidate input sources. Then, through re-ranking, one to two sources are selected and input into the question-answering model to generate answers.
- ImplicitDec [4]: Similar to AutoRouting, this model calls different sub-models for different information source modalities but generates answers to multi-hop questions in a stepwise manner. It uses a question-type classifier to determine the modality of the question-related information sources, the order of the modalities, and the logical operations required to be called. During each hop, the corresponding sub-model generates an answer based on the question, the information source of the current modality, and the answers generated in previous hops. The sub-models applied for each modality are similar to those used in AutoRouting.
- MuRAG [6]: This model is pre-trained on a large-scale corpus of image–text pairs and pure text. It retrieves the top-k nearest neighbors from the memory of image–text pairs

based on the question. The retrieved results and the question are concatenated and input into an encoder–decoder to generate an answer. The model uses ViT to encode images and T5 [35] to encode text and generate answers.

4.4. Evaluation Metrics

Due to the difference in answer formats within the MultimodalQA dataset (text fragments or phrases) and the WebQA dataset (natural language sentences), different evaluation metrics are used for the two datasets. For the MultimodalQA dataset, the evaluation metrics include Exact Match (EM) and F1. For the WebQA dataset, the evaluation metrics include retrieval results F1, QA-FL, QA-Acc, and QA-Overall. Exact Match (EM) is used to assess whether the generated answer is completely identical to the reference answer. The EM value is 1 only if the two are exactly the same; otherwise, it is 0. The F1 score is the harmonic mean of precision and recall between the candidate answer and the standard answer. QA-FL measures the fluency (syntactic and semantic relevance) between the generated answer and the reference answer, with its core being BARTScore [36]. BARTScore is a natural language generation evaluation metric based on paraphrase quality measurement, that is, evaluating the quality of generated text by the probability of text generation. BARTScore(A,B) indicates the probability of generating text B from text A. Given a reference answer r and a generated candidate answer c, BARTScore(r,c) represents the probability of generating the candidate answer from the reference answer. QA-FL prioritizes semantic consistency and is somewhat robust to the misplacement of function words. It penalizes the rearrangement of words and expressions of disfluency. Considering the presence of multiple reference answers, the QA-FL score for the generated candidate answer c is represented as follows:

$$FL(c, R) = max\left\{min(1, (\frac{BARTScore(r, c)}{BARTScore(r, r)}))\right\} \tag{11}$$

QA-Acc is used to measure the overlap of key entities between the generated answer and the reference answer. To ensure the accuracy of the answer, it is necessary to ensure the presence of specific key entities in the answer and penalize any incorrect entities present. For different question categories (qc), different answer domains (D_{qc}) are defined, which are the sets of keywords related to the category. Given a candidate answer c and reference answer keywords K, the QA-Acc metric is calculated as follows:

$$Acc(c, K) = \begin{cases} F1(c \cap D_{qc}, K \cap D_{qc}) & qc \in [color, shape, number, Y/N] \\ Recall(c, K) & otherwise: \end{cases} \tag{12}$$

QA-Overall is used to measure the model's overall performance for the dataset. It is calculated as the average of the product of QA-FL and QA-Acc for each sample across the entire test set.

4.5. Results

The results of our model (PEG) and baseline models for the WebQA dataset are shown in Table 1. Compared to baseline models, our model achieves leading results across all metrics. For the Retrieval-F1, our model outperforms the MuRAG model by 3.8. For the QA-FL, which measures the fluency of the answers, our model exceeds the MuRAG model by 0.4. For the QA-Acc, assessing the accuracy of the answers, our model surpasses the MuRAG model by 1.9. For the QA-Overall, which represents the overall score of the answers, our model leads the MuRAG model by 1.7. This indicates that the PEG model demonstrates superior performance in both retrieval effectiveness and the quality of generated answers, highlighting its effectiveness in multimodal question answering tasks for the WebQA dataset.

Table 1. Experiment results for WebQA dataset.

Model	Retrieval-F1	QA-FL	QA-Acc	QA-Overall
Question-Only	-	34.9	22.2	13.4
VLP (Oscar)	68.9	42.6	36.7	22.6
VLP + ResNeXt	69.0	43.0	37.0	23.0
VLP + VinVL	70.9	44.2	38.9	24.1
MuRAG	74.6	55.7	54.6	36.1
PEG	79.4	56.1	56.7	37.8

The results of our model and baseline models for the MultimodalQA dataset are shown in Table 2. Compared to the baseline models, our model achieves the best results across all metrics. For text questions, our model shows a significant improvement over the previously best-performing MuRAG model, with an increase of 3.9 for the F1 and an increase of 3.0 for the EM. For image questions, the proposed model shows an improvement of 2.1 for the F1 over the MuRAG model.

Table 2. Experiment results for MultimodalQA dataset.

Model	Text		Image		All
	EM	F1	EM	F1	EM
Question-Only	15.4	18.4	11.0	15.6	13.8
AutoRouting	49.5	56.9	37.8	37.8	46.6
MuRAG	60.8	67.5	58.2	58.2	60.2
PEG	64.7	70.5	60.3	60.3	63.8

The experimental results for both datasets indicate that our model outperforms baseline models. Generating supplementary descriptions for images allows the model to better understand and align images and text, thereby achieving better retrieval results. By generating background knowledge for questions and pairing it with more accurate input information sources, the model can generate higher quality answers.

4.6. Ablation Study

To better analyze the effectiveness of each module within our proposed model, we conducted ablation experiments by removing one of the three key parts at a time for ablation analysis. PEG w/o IC: By removing the image supplementary description generation module, the texts within the image–text pairs of the input information sources are no longer expanded. During information source retrieval and answer generation, the textual content in the relevant information sources remains as the original image descriptions, without any supplementary descriptions. PEG w/o CON: During the training process of the PEG model, answers are generated directly using question-related information sources, and the contrastive learning loss is removed. PEG w/o BK: The background knowledge generation module is removed. When generating answers, the model only inputs the question and relevant information sources, without including background knowledge related to the question.

Apart from the modifications to the aforementioned modules, all other modules and experimental settings are consistent with the PEG model. Tables 3 and 4 present the results of ablation experiments for the WebQA and MultimodalQA datasets, respectively.

Comparing the two tables, it is evident that the model performance significantly decreases after removing the contrastive learning component. With the WebQA dataset, the Retrieval-F1 metric drops by 6.3, QA-FL by 1.2, QA-Acc by 2.6, and QA-Overall by 2.4. With the MultimodalQA dataset, for text-related questions, F1 and EM decrease by 4.5 and 3.5, respectively, and for image-related questions, F1 and EM each decrease by 2.2.

This decline is attributed to the crucial role of contrastive learning in the vision–language joint encoding process, which promotes the encoder's alignment and understanding of images and texts. The encoder's capability directly affects the quality of information source retrieval and thereby the quality of answer generation.

Table 3. Ablation study results for WebQA dataset.

Model	Retrieval-F1	QA-FL	QA-Acc	QA-Overall
PEG	79.4	56.1	56.7	37.8
PEG w/o IC	76.0	55.7	55.3	36.3
PEG w/o CON	73.1	54.9	54.1	35.4
PEG w/o BK	78.8	55.9	55.8	36.8

Table 4. Ablation study results for MultimodalQA dataset.

Model	Text		Image		All
	EM	F1	EM	F1	EM
PEG	64.7	70.5	60.3	60.3	63.8
PEG w/o IC	62.4	68.9	59.5	59.5	61.7
PEG w/o CON	60.2	67.0	58.1	58.1	60.1
PEG w/o BK	63.3	69.6	59.8	59.8	62.5

After removing the image supplementary description generation, for the WebQA dataset, the Retrieval-F1 metric decreased by 3.4, QA-FL by 0.4, QA-Acc by 1.4, and QA-Overall by 1.5. For the MultimodalQA dataset, for text-related questions, F1 and EM decreased by 2.3 and 1.6, respectively, and for image-related questions, F1 and EM each decreased by 0.8. With the removal of supplementary descriptions, the vision–language joint encoder experiences a reduction in text information, which in turn diminishes its encoding capability, leading to a decrease in retrieval performance. Additionally, the reduction in input text information weakens the model's ability to generate answers, as the encoder has less contextual data to work with, impacting the overall effectiveness of the model in generating accurate and relevant answers.

After removing the background knowledge generation, for the WebQA dataset, the Retrieval-F1 metric decreased by 0.6, QA-FL by 0.2, QA-Acc by 0.9, and QA-Overall by 0.8. For the MultimodalQA dataset, for text-related questions, F1 and EM decreased by 1.4 and 0.9, respectively, and for image-related questions, F1 and EM each decreased by 0.5. Removing background knowledge results in a reduction in input for the vision–language joint encoder, slightly diminishing its encoding capability, which leads to a minor decrease in retrieval performance. Additionally, reducing the input of background knowledge causes a decline in the model's ability to generate answers.

The results of the ablation experiments demonstrate that the image supplementary description generation, contrastive learning in information source encoding, and background knowledge generation modules included in the proposed model significantly enhance the model's capability to retrieve information sources and generate answers effectively.

To analyze the impact of different prompt learning methods, we replaced prefix-tuning with P-tuning [37] and P-tuning v2 [38] in background knowledge generation. The experimental results are shown in Tables 5 and 6. PEG: Our proposed model includes all modules and utilizes prefix-tuning to generate background knowledge. PEG w/o BK: The background knowledge generation module is removed. When generating answers, the model only inputs the question and relevant information sources, without including background knowledge related to the question. PEG with P-tuning: In the PEG model, we replace prefix-tuning with P-tuning to generate background knowledge, while keeping the other modules unchanged. PEG with P-tuning v2: In the PEG model, we replace

prefix-tuning with P-tuning v2 to generate background knowledge, while keeping the other modules unchanged.

Table 5. Results of Different prompt learning methods for WebQA dataset.

Model	Retrieval-F1	QA-FL	QA-Acc	QA-Overall
PEG	79.4	56.1	56.7	37.8
PEG w/o BK	78.8	55.9	55.8	36.8
PEG with P-tuning	79.1	56.0	56.2	37.1
PEG with P-tuning v2	79.8	56.4	57.1	38.1

Table 6. Results of Different prompt learning methods for MultimodalQA dataset.

Model	Text		Image		All
	EM	F1	EM	F1	EM
PEG	64.7	70.5	60.3	60.3	63.8
PEG w/o BK	63.3	69.6	59.8	59.8	62.5
PEG with P-tuning	63.7	70.0	59.9	59.9	63.0
PEG with P-tuning v2	65.2	70.9	60.5	60.5	64.2

From the experimental results, it is evident that using any prompt learning method is better than not using any prompt learning method to generate background knowledge. The effect of using P-tuning v2 is better than using prefix-tuning and P-tuning. However, the effect of using prefix-tuning is better than P-tuning. This is because P-tuning introduces learnable parameters in the input without changing the main weights of the language model, and it adjusts the minimal number of parameters. On the other hand, prefix-tuning optimizes prefix parameters for all layers. P-tuning v2, based on P-tuning, incorporates the idea of prefix-tuning by inserting continuous prompts into each layer, with prompts between layers being independent of each other. It can be seen as an optimized version of prefix-tuning.

To verify the impact of the number of information sources retrieved (top-k) on the retrieval results and the quality of answer generation, Figure 6 illustrates how the retrieval performance metric Retrieval-F1 and the answer generation performance metric QA-Overall change as the number of retrieved information sources (top-k) varies. In the WebQA dataset, the number of information sources related to a question is mostly more than two but rarely exceeds four. Therefore, as top-k gradually increases from 2, both Retrieval-F1 and QA-Overall metrics improve. However, when top-k exceeds 4, the input of excessive information sources introduces noise, leading to a decline in both Retrieval-F1 and QA-Overall metrics.

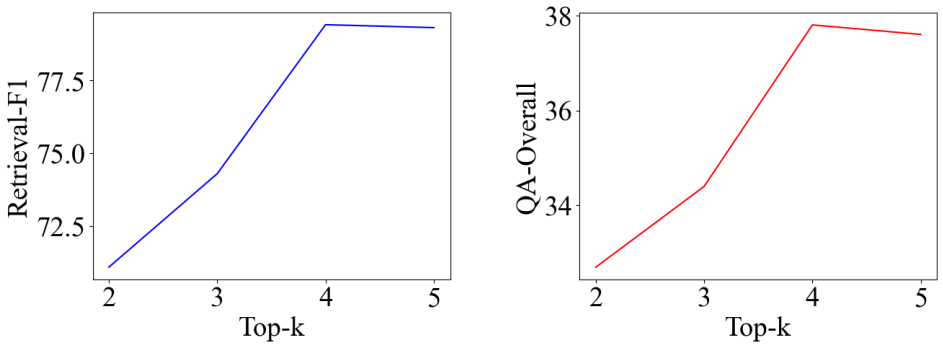

Figure 6. The impact of top-k on retrieval and answer generation.

4.7. Case Study

Figure 7 presents an example of an answer generated by our proposed model. It can be observed that the original textual description in the information source provides the location of the image but does not describe any content within the image. The model, through image supplementary description generation, broadly describes the image content (building, cars, bikes). The correspondence between the image and text significantly aids the vision–language joint encoder in aligning and understanding the image content. The answer generated by our model is essentially consistent with the reference answer, proving the effectiveness of the model.

Quetion: Is the fence in front of The Glass House in Fulham taller or shorter than a bicycle?

Description: The Glass House, Fulham 05

Supplementary Description： It occupies a brick building graced with three large windows trimmed in white. Cars park neatly on the street ahead, bikes rest against railings closeby.

Generated answer: The fence in front of the Glass House appears to be slightly taller than a bicycle.
.
Ground truth: The fence in font of the building is taller than a typical bicycle.

Figure 7. Examples of PEG-generated answers.

Table 7 shows example of a text-only question in the WebQA dataset. In this case, the relevant input information sources for the given questions do not contain images, only textual content. The text related to the question retrieved by the model, Text 1, is inaccurate and does not contain knowledge relevant to the question; it only has keyword overlap. However, the generated background knowledge provides information related to the question, which is beneficial for the model to generate the correct answer.

Table 7. Example of PEG-generated answer to text-only question.

Question:	**What Body of Water Is Both Trinidad and the Petites Antilles in?**
Relatex text 1:	Major bodies of water on Trinidad include the Hollis Reservoir, Navet Reservoir, Caroni Reservoir. Trinidad is made up of a variety of soil types, the majority being fine sands and heavy clays.
Relatex text 2:	The Lesser Antilles (Spanish: Pequeas Antillas; French: Petites Antilles; Papiamento: Antias Menor; Dutch: Kleine Antillen) is a group of islands in the Caribbean Sea. Most form a long, partly volcanic island arc between the Greater Antilles to the north-west and the continent of South America.
Background knowledge:	Trinidad is situated in the Caribbean Sea, which is a part of the western Atlantic Ocean.
Genarated answer:	Both Trinidad and the Petites Antilles are in the Caribbean Sea.
Ground truth:	The Caribbean.

5. Conclusions

In this paper, we propose a prompt-enhanced generation model for multimodal question answering. To address the issue of insufficient image descriptions in datasets, which leads to inadequate understanding and alignment between images and texts, we utilize a multimodal pre-trained model to generate supplementary captions for images. By employing a vision–language joint encoder that fully interacts and aligns images with texts, we

improve the encoding effectiveness, which in turn enhances retrieval capabilities. Since the quality of generated answers is closely related to the ability to retrieve relevant information sources, inputting unrelated information sources into the model can prevent it from correctly answering questions. To solve this issue, we adopt a prefix-tuning approach to mine knowledge present in the parameters of pre-trained language models, generating related background knowledge for the question as a supplement to the information sources for retrieval. We conducted experiments and analysis on two datasets to verify the effectiveness of our model and its related modules.

For future work, we will focus on more efficiently integrating prompt learning with retrieval to minimize the computational cost of retrieval as much as possible, avoid the hallucinations brought about by model prompt learning, and retrieve relevant information when necessary to enhance the accuracy of generated answers.

Author Contributions: C.C.: Conceptualization, methodology, validation, investigation, writing—original draft preparation; Z.L.: writing—review and editing, supervision. All authors have read and agreed to the published version of the manuscript.

Funding: This work was supported in part by the National Natural Science Foundation of China (Grant Nos. 62276017, U1636211, 61672081), and the Fund of the State Key Laboratory of Software Development Environment (Grant No. SKLSDE-2021ZX-18).

Data Availability Statement: The WebQA dataset is available at https://webqna.github.io/ (accessed on 9 April 2024); The MultimodalQA dataset is available at https://allenai.github.io/multimodalqa/ (accessed on 9 April 2024). Our code will be available at https://github.com/libracui/PEG (accessed on 9 April 2024).

Conflicts of Interest: The authors declare no conflicts of interest.

References

1. Hannan, D.; Jain, A.; Bansal, M. Manymodalqa: Modality disambiguation and qa over diverse inputs. In Proceedings of the AAAI Conference on Artificial Intelligence, New York, NY, USA, 7–12 February 2020; pp. 7879–7886.
2. Reddy, R.G.; Rui, X.; Li, M.; Lin, X.; Wen, H.; Cho, J.; Huang, L.; Bansal, M.; Sil, A.; Chang, S.F.; et al. Mumuqa: Multimedia multi-hop news question answering via cross-media knowledge extraction and grounding. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtually, 22 February–1 March 2022; pp. 11200–11208.
3. Chang, Y.; Narang, M.; Suzuki, H.; Cao, G.; Gao, J.; Bisk, Y. WebQA: Multihop and Multimodal QA. In Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), New Orleans, LA, USA, 18–24 June 2022; pp. 16495–16504.
4. Talmor, A.; Yoran, O.; Catav, A.; Lahav, D.; Wang, Y.; Asai, A.; Ilharco, G.; Hajishirzi, H.; Berant, J. MultiModalQA: complex question answering over text, tables and images. In Proceedings of the International Conference on Learning Representations, Vienna, Austria, 4 May 2021.
5. Zhou, L.; Palangi, H.; Zhang, L.; Hu, H.; Corso, J.; Gao, J. Unified Vision-Language Pre-Training for Image Captioning and VQA. *Proc. Aaai Conf. Artif. Intell.* **2020**, *34*, 13041–13049. [CrossRef]
6. Chen, W.; Hu, H.; Chen, X.; Verga, P.; Cohen, W. MuRAG: Multimodal Retrieval-Augmented Generator for Open Question Answering over Images and Text. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, Abu Dhabi, United Arab Emirates, 7–11 December 2022; Goldberg, Y., Kozareva, Z., Zhang, Y., Eds.; Association for Computational Linguistics: Abu Dhabi , United Arab Emirates, 2022; pp. 5558–5570.
7. Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; Sutskever, I. Language models are unsupervised multitask learners. *Openai Blog* **2019**, *1*, 9.
8. Wang, C.; Liu, P.; Zhang, Y. Can Generative Pre-trained Language Models Serve As Knowledge Bases for Closed-book QA? In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Virtual Event, 1–6 August 2021; pp. 3241–3251.
9. Brown, T.; Mann, B.; Ryder, N.; Subbiah, M.; Kaplan, J.D.; Dhariwal, P.; Neelakantan, A.; Shyam, P.; Sastry, G.; Askell, A.; et al. Language models are few-shot learners. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 1877–1901.
10. Mahabadi, R.K.; Zettlemoyer, L.; Henderson, J.; Mathias, L.; Saeidi, M.; Stoyanov, V.; Yazdani, M. Prompt-free and Efficient Few-shot Learning with Language Models. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Dublin, Ireland, 22–27 May 2022; pp. 3638–3652.
11. Antol, S.; Agrawal, A.; Lu, J.; Mitchell, M.; Batra, D.; Zitnick, C.L.; Parikh, D. Vqa: Visual question answering. In Proceedings of the International Conference on Computer Vision, Santiago, Chile, 7–13 December 2015; pp. 2425–2433.

12. Marino, K.; Rastegari, M.; Farhadi, A.; Mottaghi, R. Ok-vqa: A visual question answering benchmark requiring external knowledge. In Proceedings of the IEEE/cvf Conference on Computer Vision and Pattern Recognition, Long Beach, CA, USA, 15–20 June 2019; pp. 3195–3204.
13. Singh, H.; Nasery, A.; Mehta, D.; Agarwal, A.; Lamba, J.; Srinivasan, B.V. Mimoqa: Multimodal input multimodal output question answering. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Online, 6–11 June 2021; pp. 5317–5332.
14. Lin, W.; Byrne, B. Retrieval Augmented Visual Question Answering with Outside Knowledge. In Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, Abu Dhabi, United Arab Emirates, 7–11 December 2022; pp. 11238–11254.
15. Yang, Z.; Gan, Z.; Wang, J.; Hu, X.; Lu, Y.; Liu, Z.; Wang, L. An empirical study of gpt-3 for few-shot knowledge-based vqa. In Proceedings of the AAAI Conference on Artificial Intelligence, Virtually, 22 February–1 March 2022; Volume 36, pp. 3081–3089.
16. Yu, B.; Fu, C.; Yu, H.; Huang, F.; Li, Y. Unified Language Representation for Question Answering over Text, Tables, and Images. In Proceedings of the Findings of the Association for Computational Linguistics: ACL 2023, Toronto, ON, Canada, 9–14 July 2023; pp. 4756–4765.
17. Shi, Q.; Cui, H.; Wang, H.; Zhu, Q.; Che, W.; Liu, T. Exploring Hybrid Question Answering via Program-based Prompting. *arXiv* **2024**, arXiv:2402.10812.
18. Karpukhin, V.; Oguz, B.; Min, S.; Lewis, P.; Wu, L.; Edunov, S.; Chen, D.; Yih, W.T. Dense Passage Retrieval for Open-Domain Question Answering. In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Online, 6–20 November 2020; pp. 6769–6781.
19. Guu, K.; Lee, K.; Tung, Z.; Pasupat, P.; Chang, M. Retrieval augmented language model pre-training. In Proceedings of the International conference on machine learning. PMLR, Virtual Event, 13–18 July 2020; pp. 3929–3938.
20. Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.t.; Rocktäschel, T.; et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Adv. Neural Inf. Process. Syst.* **2020**, *33*, 9459–9474.
21. Izacard, G.; Grave, É. Leveraging Passage Retrieval with Generative Models for Open Domain Question Answering. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, Online, 19–23 April 2021; pp. 874–880.
22. Wang, P.; Yang, A.; Men, R.; Lin, J.; Bai, S.; Li, Z.; Ma, J.; Zhou, C.; Zhou, J.; Yang, H. OFA: Unifying Architectures, Tasks, and Modalities Through a Simple Sequence-to-Sequence Learning Framework. In Proceedings of the 39th International Conference on Machine Learning, Baltimore, MD, USA, 17–23 July 2022; Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., Sabato, S., Eds.; PMLR (Proceedings of Machine Learning Research): Baltimore, ML, USA, 2022; Volume 162, pp. 23318–23340.
23. Dosovitskiy, A.; Beyer, L.; Kolesnikov, A.; Weissenborn, D.; Zhai, X.; Unterthiner, T.; Dehghani, M.; Minderer, M.; Heigold, G.; Gelly, S.; et al. An Image is Worth 16 $\times$ 16 Words: Transformers for Image Recognition at Scale. In Proceedings of the International Conference on Learning Representations, Vienna, Austria, 4 May 2021.
24. Guo, R.; Sun, P.; Lindgren, E.; Geng, Q.; Simcha, D.; Chern, F.; Kumar, S. Accelerating Large-Scale Inference with Anisotropic Vector Quantization. In Proceedings of the 37th International Conference on Machine Learning, Virtual Event, 13–18 July 2020; III, H.D., Singh, A., Eds.; PMLR (Proceedings of Machine Learning Research): Baltimore, ML, USA, 2020; Volume 119; pp. 3887–3896.
25. Roberts, A.; Raffel, C.; Shazeer, N. How Much Knowledge Can You Pack Into the Parameters of a Language Model? In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Online, 16–20 November 2020; pp. 5418–5426.
26. Li, X.L.; Liang, P. Prefix-Tuning: Optimizing Continuous Prompts for Generation. In Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), Online, 1–6 August 2021 ; pp. 4582–4597.
27. Lewis, M.; Liu, Y.; Goyal, N.; Ghazvininejad, M.; Mohamed, A.; Levy, O.; Stoyanov, V.; Zettlemoyer, L. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In Proceedings of the ACL 2020, Online, 5–10 July 2020; pp. 7871–7880.
28. Kingma, D.P.; Ba, J. Adam: A method for stochastic optimization. In Proceedings of the ICLR 2015, San Diego, CA, USA, 7–9 May 2015.
29. Li, X.; Yin, X.; Li, C.; Zhang, P.; Hu, X.; Zhang, L.; Wang, L.; Hu, H.; Dong, L.; Wei, F.; et al. Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks. In Proceedings of the Computer Vision—ECCV 2020: 16th European Conference, Glasgow, UK, 23–28 August 2020; pp. 121–137.
30. Zhang, P.; Li, X.; Hu, X.; Yang, J.; Zhang, L.; Wang, L.; Choi, Y.; Gao, J. VinVL: Revisiting Visual Representations in Vision-Language Models. In Proceedings of the 2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR), Nashville, TN, USA, 20–25 June 2021.
31. Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; Stoyanov, V. RoBERTa: A Robustly Optimized BERT Pretraining Approach. 2019. Available online: https://arxiv.org/abs/1907.11692 (accessed on 9 April 2024).
32. Lu, J.; Batra, D.; Parikh, D.; Lee, S. ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks. *Neural Inf. Process. Syst. Inf. Process. Syst.* **2019**, *32* , 13–23.
33. Ren, S.; He, K.; Girshick, R.; Sun, J. Faster R-CNN: Towards Real-Time Object Detection with Region Proposal Networks. *IEEE Trans. Pattern Anal. Mach. Intell.* **2017**, *39* , 1137–1149. [CrossRef] [PubMed]

34. Radford, A.; Kim, J.W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. Learning Transferable Visual Models From Natural Language Supervision. In Proceedings of the 38th International Conference on Machine Learning, Virtual Event, 18–24 July 2021; Meila, M., Zhang, T., Eds.; PMLR (Proceedings of Machine Learning Research): Virtual, 2021; Volume 139, pp. 8748–8763.
35. Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; Liu, P.J. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.* **2020**, *21* , 1–67.
36. Yuan, W.; Neubig, G.; Liu, P. BARTScore: Evaluating Generated Text as Text Generation. In Proceedings of the Advances in Neural Information Processing Systems, Denver, CO, USA, 6–14 December 2021; Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., Vaughan, J.W., Eds.; Curran Associates, Inc.: Virtual, 2021; Volume 34, pp. 27263–27277.
37. Liu, X.; Zheng, Y.; Du, Z.; Ding, M.; Qian, Y.; Yang, Z.; Tang, J. GPT Understands, Too. *arXiv* **2021**, arXiv:2103.10385.
38. Liu, X.; Ji, K.; Fu, Y.; Tam, W.; Du, Z.; Yang, Z.; Tang, J. P-Tuning: Prompt Tuning Can Be Comparable to Fine-tuning Across Scales and Tasks. In Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Dublin, Ireland, 22–27 May 2022; pp. 61–68. [CrossRef]