Multimodal Reranking for Knowledge-Intensive Visual Question Answering

Haoyang Wen[*§]
Carnegie Mellon University
hwen3@cs.cmu.edu

Honglei Zhuang
Google
hlz@google.com

Hamed Zamani[*]
University of Massachusetts Amherst
zamani@cs.umass.edu

Alexander Hauptmann
Carnegie Mellon University
alex@cs.cmu.edu

Michael Bendersky
Google
bemike@google.com

Abstract

Knowledge-intensive visual question answering requires models to effectively use external knowledge to help answer visual questions. A typical pipeline includes a knowledge retriever and an answer generator. However, a retriever that utilizes local information, such as an image patch, may not provide reliable question-candidate relevance scores. Besides, the two-tower architecture also limits the relevance score modeling of a retriever to select top candidates for answer generator reasoning. In this paper, we introduce an additional module, a multi-modal reranker, to improve the ranking quality of knowledge candidates for answer generation. Our reranking module takes multi-modal information from both candidates and questions and performs cross-item interaction for better relevance score modeling. Experiments on OK-VQA and A-OKVQA show that multi-modal reranker from distant supervision provides consistent improvements. We also find a training-testing discrepancy with reranking in answer generation, where performance improves if training knowledge candidates are similar to or noisier than those used in testing.

1 Introduction

Knowledge-intensive visual question answering (KI-VQA), compared to conventional visual question answering, provides questions that cannot be directly answered with images. It requires models to use external knowledge for answer reasoning and synthesis, as shown in Figure 1.

Q: What US city is associated with this type of pizza?
A: Chicago

Figure 1: An example from OK-VQA, which requires knowledge to associate deep-dish pizza and Chicago.

A typical KI-VQA system contains a retrieval model to find relevant external knowledge, and an answer generator that performs reasoning over retrieved knowledge to produce the answer. One line of research investigates methods for an effective retrieval pipeline, which includes the choices of knowledge bases (Li et al., 2020; Gardères et al., 2020; Luo et al., 2021), and methods on retrieval with visual descriptions (Luo et al., 2021) or image-text retrieval (Gui et al., 2022; Lin et al., 2022).

Answer generation models usually use retrieval relevance scores to select top candidates (Gui et al., 2022; Lin et al., 2022). Although achieving great success, it may sometimes provide unreliable scores, especially for retrieval using images. Because we usually split an image into a series of image patches and perform retrieval with individual patches, a high relevance score of one patch may not necessarily translate to a high overall question-candidate relevance. Besides, the two-tower architecture of a retriever model also lacks cross-item modeling for predicting precise relevance scores.

In this work, we propose to include multi-modal reranking to improve the relevance score modeling, as reranking have already shown its importance in various knowledge-intensive tasks (Liu, 2009; Lee et al., 2018; Wang et al., 2018; Mao et al., 2021; Glass et al., 2022; Hofstätter et al., 2023). The multi-modal reranking uses the multi-modal question and multi-modal knowledge items to obtain the relevance score. Specifically, we finetune a pretrained multi-modal language model (Chen et al., 2023b) to perform a multi-modal cross-item inter-

[*]Work performed while at Google.

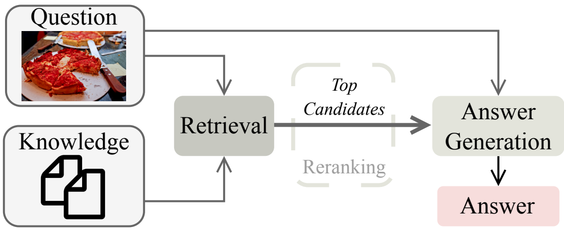

Figure 2: A basic KI-VQA framework, which first retrieves relevant top knowledge candidates with using visual question and then combine the question and retrieved knowledge candidates to generate the answer. The dashed box is our reranking module in Section 3.

action between the question and knowledge items. We train our reranker on the same dataset as answer generator training, distantly supervised by checking if answer candidates appear in the knowledge text. The benefits of this reranking component are two-folded. On one side, as other typical reranking components, it can provide more reliable relevance scores by modeling the cross-item interaction. On the other side, because most of the existing retrieval models performs uni-modal retrieval (Luo et al., 2021; Gui et al., 2022; Lin et al., 2022), reranking with multi-modal interaction can improve the quality of retrieval by multi-modal information from question and knowledge candidates.

We perform experiments on OK-VQA (Marino et al., 2019) and A-OKQVA (Schwenk et al., 2022), based on image-text retrieval (Jia et al., 2021). The results show that the distantly-supervised reranker provides consistent improvement compared to the pipeline without a reranker. We also observe a training-testing discrepancy with reranking for answer generation, finding that performance improves when training knowledge candidates are similar to or noisier than testing candidates. We also find that an oracle reranker can provide a promising performance upperbound, which sheds light on future research directions in this area.

2 A Knowledge-Intensive Visual Question Answering Framework

In this section, we will introduce a basic framework for KI-VQA, including image-text retrieval and answer generation, as illustrated in Figure 2.

2.1 Wikipedia-Based Image Text Dataset

In this work, we use a multi-modal knowledge base, Wikipedia-Based Image Text Dataset (WIT) (Srinivasan et al., 2021). In addition to previous work that uses text from encyclopedia, WIT contains images from Wikipedia and the surrounding text at different levels, including their captions and surrounding sections. Therefore, we consider WIT as a combination of image and text knowledge.

2.2 Image-Text Retrieval

Previous work has explored the use of different retrieval model choices (Luo et al., 2021; Gui et al., 2022; Lin et al., 2022). We follow one line of research that adopts image-text retrieval (Gui et al., 2022) using pretrained image-text language model with dual-encoder architecture (Radford et al., 2021; Jia et al., 2021). Following Gui et al. (2022), we use sliding window with a stride to generate multiple image regions from question image. Each image region is considered as a query and will be encoded by image encoder model $\phi_i(\cdot)$. We encode captions in WIT dataset as the representation for candidates with text encoder model $\phi_t(\cdot)$, as captions in Wikipedia are generally informative. Relevance score between an image region v_i and a WIT candidate c is obtained with the inner product of their representations

$$r_t(v_i, c) = \phi_i(v_i)^T \phi_t(c).$$

2.3 Answer Generation

We follow previous work (Gui et al., 2022; Lin et al., 2022) that performs reasoning over top candidates within an encoder-decoder architecture. We also incorporate the multi-modal information (Salemi et al., 2023), compared to previous work that mostly uses text-based information.

Our answer generation module is finetuned on vision language models that takes the combination of image and text as input (*e.g.*, Chen et al., 2023b; Li et al., 2023). We first encode each top candidate separately. The input of each candidate consists of question image, candidate image and text following a template[1] to compose question and candidate. We encode the image with a Vision Transformer (Dosovitskiy et al., 2021), which takes a series of image patches $\boldsymbol{x}^v = [x^v_1, \ldots, x^v_n]$, *i.e.*, image tokens, to produce image representations

$$\boldsymbol{E}^v = [\boldsymbol{e}^v_1, \ldots, \boldsymbol{e}^v_n] = Enc_v(\boldsymbol{x}^v).$$

We combine image representations and text token embeddings $\boldsymbol{E}^t$ to produce fused representations with a Transformer (Vaswani et al., 2017)

$$\boldsymbol{H} = [\boldsymbol{H}^v_q; \boldsymbol{H}^v_c; \boldsymbol{H}^t] = Enc_t\left([\boldsymbol{E}^v_q; \boldsymbol{E}^v_c; \boldsymbol{E}^t]\right),$$

[1] `question: <question text> candidate: <caption>`

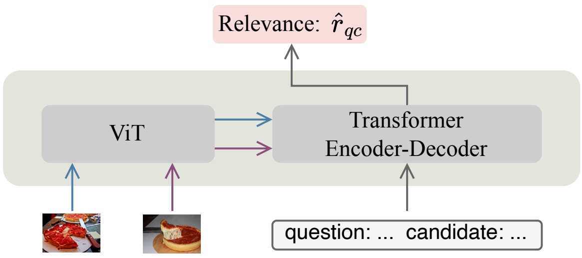

Figure 3: Framework of multimodal reranking.

where $\boldsymbol{E}_q^v$, $\boldsymbol{H}_c^v$ represent the image token representations for question and candidate image, respectively. We also include an empty candidate that consists only of the question image and text.

While decoding, to reduce the total number of representations, we only keep the question image and text representations from the empty candidate, and the token representations that corresponds to each knowledge caption text. We concatenate these token representations to form a global representation for decoder to perform cross-attention and generate each answer token autoregressively (Izacard and Grave, 2021b).

3 Multi-Modal Reranking

Vanilla retrieval-generation frameworks directly use relevance score from individual image patches. However, a high relevance from a region does not necessarily imply the overall relevance. In this section, we propose multi-modal reranking, as illustrated in Figure 3, which takes multi-modal question and knowledge as input and produces relevance scores with cross-item interaction.

3.1 Modeling

Our ranking model is also finetuned on the multi-modal pretrained language model. For each question-candidate pair, we first encode the question and candidate image separately, and obtain two series of token representations $\boldsymbol{E}_q^v, \boldsymbol{E}_c^v$. Then we concatenate the two series of image token representations, with text token embeddings $\boldsymbol{E}^t$ following same template in Section 2.3 for a Transformer to produce fused token representations

$$\boldsymbol{H}^r = Enc_r\left(\left[\boldsymbol{E}_q^v; \boldsymbol{E}_c^v; \boldsymbol{E}_q^t; \boldsymbol{E}_t^t\right]\right).$$

We follow Zhuang et al. (2023) and use 1-step decoding to obtain the score from the unnormalized loglikelihood of a special token “<extra_id_10>”

$$\hat{r}_{qc} = Dense\left(Dec\left(\boldsymbol{H}^r\right)\right)_{(\texttt{<extra\_id\_10>})}.$$

3.2 Ranker Training

Because we do not have ground-truth relevance scores, we adopt distant supervision labels for reranking training. In a typical VQA setting, each answer consists of 10 answer candidate annotations. We count the number of answer candidates that occur in the knowledge candidate text as o. The distantly supervised relevance score is obtained similar to VQA accuracy (Antol et al., 2015)

$$r_{qc} = \min\left\{o/3, 1\right\}.$$

On OK-VQA, we split the training dataset of original dataset into sub-training and -development sets. At each training step, for a question q, we uniformly sample a candidate set $\mathbb{C}$ from the retrieval results, and apply pairwise logistics ranking loss (Burges et al., 2005), which compares the ranking between all pairs of candidates in the set

$$\ell(q) = \sum_{c\in\mathbb{C}}\sum_{c'\in\mathbb{C}} \mathbb{I}_{r_{qc}>r_{qc'}} \log\left(1 + e^{\hat{r}_{qc'} - r_{qc}}\right).$$

3.3 Discrepancy on Applying Reranking for Answer Generation

During answer generation training, it is straightforward to apply the ranking model and use the reranked top candidates as input. However, directly applying reranking on both training and testing will instead hurt the model performance. This is because applying the ranker on the training set, from which the ranker is trained, performs much better than when applied to the unseen test set. As we will illustrate in Section 4.3, learning answer generation with higher quality ranking results while testing on lower quality ranking results will in general have a negative impact to answer generation performance. Therefore, we will keep the initial retrieval results for answer generation training, while using the reranked results for model testing.

4 Experiments

4.1 Setup

We conduct experiments on OK-VQA (Marino et al., 2019) and A-OKVQA (Schwenk et al., 2022). OK-VQA introduces visual questions that requires external knowledge. A-OKVQA further emphasizes commonsense reasoning over world knowledge. For both datasets, we evaluate the performance on the validation set. Following the standard setting, we use VQA accuracy as the metric.

Methods	VQA Accuracy
BAN+KG (Li et al., 2020)	26.7
Mucko (Zhu et al., 2020)	29.2
ConceptBERT (Gardères et al., 2020)	33.7
KRISP (Marino et al., 2021)	38.9
Vis-DPR (Luo et al., 2021)	39.2
MAVEx (Wu et al., 2022)	40.3
KAT (Gui et al., 2022)	44.3
TRiG (Gao et al., 2022)	49.4
Our model	**52.6**
models with GPT-3 generated candidates	
PICa (Yang et al., 2022)	48.0
KAT (Gui et al., 2022)	53.1
REVIVE (Lin et al., 2022)	56.6
Our model + REVIVE GPT-3	**57.2**
Our model w/ oracle ranking	64.4

Table 1: Results comparison on OK-VQA dataset.

Methods	VQA Accuracy
ViLBERT (Lu et al., 2019)	30.6
LXMERT (Tan and Bansal, 2019)	30.7
KRISP (Marino et al., 2021)	33.7
GPV-2 (Kamath et al., 2022)	48.6
Our model	**51.6**

Table 2: Results comparision on A-OKVQA dataset.

We use ALIGN (Radford et al., 2021) for image-text retrieval, and use PaLI (Chen et al., 2023b) to initialize (vision and text) Transformers in answer generation and reranking independently. Besides retrieved knowledge candidates, we also follow REVIVE (Lin et al., 2022) and use candidates generated from GPT-3. For our model with REVIVE GPT-3, we replace the last 5 candidates of the aggregated candidates with top-5 GPT-3 generated candidates from Lin et al. (2022). Appendix A includes a detailed experimental setup.

4.2 Results

Our results on Table 1 and Table 2 illustrate the performance compared to some existing work. Results on Table 1 show that we provide competitive performance compared to these systems. We also include a comparison for models with GPT-3 (Brown et al., 2020) generated candidates. We find that our framework can further improve the answer generation quality with GPT-3 generated candidates from Lin et al. (2022) and outperform these baselines.

We also show that an oracle ranking from distant supervision can provide a promising upper bound, indicating that there is still a large room for future work on improving ranking in this challenge.

Methods	VQA Accuracy OK-VQA	VQA Accuracy A-OKVQA
No retrieval	50.6	50.4
+ Image Retrieval	52.1	50.3
+ Multimodal Reranking	**52.6**	**51.6**

Table 3: Effects of multimodal reranking, compared to model without retrieval and model without reranking.

Source of Candidates Train	Test	Discrepancy	VQA Accuracy
Retrieval	Retrieval	→	52.1
Reranking	Reranking	↘	50.7
Retrieval	Reranking	↗	52.6
Oracle	Oracle	→	64.4
Oracle	Retrieval	↘	47.2
Retrieval	Oracle	↗	59.7
Retrieval	Retrieval	→	52.1

Table 4: Effects of discrepancy between knowledge candidates for training and testing. → means the qualities of knowledge candidates in training and test are similar. ↘ means the quality in training is better than test. ↗ means the quality in test is better than training.

Effects of Ranking Methods. We further conduct experiments with different ranking methods to illustrate the performance of multi-modal ranking. The results are shown in Table 3. We compared variants of our model, including the model that generates answer directly without external knowledge, and the model with initial image retrieval without further reranking. We can find the steady improvement brought by multi-modal reranking on both datasets. We provide additional comparison to other reranking strategies in Appendix B and zero-shot multi-modal large models in Appendix C that are not instruction tuned on OK-VQA.

4.3 Training and Testing Discrepancy

As we discussed in Section 3.3, directly applying a trained ranking model on both training and testing will hurt the performance. We further illustrate it empirically in Table 4. We can find that if the model is trained on higher quality candidates while applied on lower quality candidates, we will observe a drastic performance drop. On the contrary, when the quality in test is better than in training, we can still find steady improvement. This phenomenon indicates that an answer generator trained with higher-quality data can not effectively conduct knowledge reasoning on noisier data, and we should therefore train the model with noisier data.

5 Related Work

A typical knowledge-intensive visual question answering model involves a knowledge retrieval to find relevant information, and answer generator to produce the answer (Li et al., 2020; Gardères et al., 2020; Luo et al., 2021; Yang et al., 2022; Gui et al., 2022; Lin et al., 2022; Salemi et al., 2023; Shao et al., 2023). Previous work on knowledge-intensive visual question answering explores knowledge bases in different modalities, such as text items (Luo et al., 2021; Gui et al., 2022), graph items (Li et al., 2020; Gardères et al., 2020), and the composition of image items and text items (Wu et al., 2022). Our work differs from previous work by involving multi-modal knowledge items as the knowledge base, where each item contains both image and text information.

There is also a line of research investigating answer reranking, where they first produce a list of answer candidates, and then rerank those candidates to obtain the most reliable answer (Marino et al., 2021; Si et al., 2021; Wu et al., 2022). Instead, the focus of our work is to first retrieve a set of knowledge candidates that can help answer generation, and then improve the quality of knowledge candidate set through multimodal knowledge candidate reranking. Those selected candidates will still serves as additional knowledge input for answer generation reasoning.

6 Conclusion

In this paper, we introduce reranking, a critical stage for knowledge-intensive tasks, into KI-VQA. Our multi-modal reranking component takes multi-modal questions and knowledge candidates as input and perform cross-item interaction. Experiments show that our proposed multi-modal reranking can provide better knowledge candidates and improve the answer generation accuracy. Experiments on the training-testing discrepancy indicate that incorporating noisier knowledge candidates during training enhances model robustness, while training with higher quality candidates than those used in testing negatively impacts performance.

Limitations

In this paper, we focus on applying multi-modal reranking to KI-VQA. However, because of the nature of visual data, directly adding visual information may significantly increase input size, and we will require more total memory to train the model. In this paper, to reduce the total memory use, we have a much smaller number of knowledge candidates for reasoning in answer generation module compared to previous work which only uses text-based knowledge candidates. Nevertheless, it is still important to further investigate more efficient ways to incorporate visual information.

Although multi-modal reranking achieves promising performance on knowledge-intensive visual question answering, it is still an open question that whether multi-modal reranking can be used help other vision-language tasks. Besides, it is also important to develop a benchmark to evaluate multi-modal reranking models systematically, which is not covered by this work.

Similarly, in this work, we only use ALIGN and PaLI as the pretrained model for retrieval, reranking and answer generation. Although it is natural to extend the framework in this work to other pretrained models, it is still interesting to see how it contributes to different (large and small) models. We provide some preliminary results comparing our reranking pipeline with zero-shot multi-modal large models (Alayrac et al., 2022; Li et al., 2023) in Appendix C, but we also notice that some work (Liu et al., 2023; Chen et al., 2023a) uses OK-VQA as instruction tuning data, making it hard to compare/be adopted directly.

We also notice that there is another line of research investigating how to effectively use large language models for knowledge-intensive visual question answering (Yang et al., 2022; Gui et al., 2022; Lin et al., 2022; Salemi et al., 2023; Shao et al., 2023). Although our preliminary results show that our framework can still provide additional improvements over same the large language model queries as in Lin et al. (2022), it is still an open question to effectively use and combine the retrieval pipeline and large language model queries.

Acknowledgment

This work was supported in part by the Google Visiting Scholar program and the Center for Intelligent Information Retrieval. Any opinions, findings and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect those of the sponsor.

References

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel

Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob L. Menick, Sebastian Borgeaud, Andy Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karén Simonyan. 2022. Flamingo: a visual language model for few-shot learning. In *NeurIPS*.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. VQA: visual question answering. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 2425–2433. IEEE Computer Society.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Christopher J. C. Burges, Tal Shaked, Erin Renshaw, Ari Lazier, Matt Deeds, Nicole Hamilton, and Gregory N. Hullender. 2005. Learning to rank using gradient descent. In *Machine Learning, Proceedings of the Twenty-Second International Conference (ICML 2005), Bonn, Germany, August 7-11, 2005*, volume 119 of *ACM International Conference Proceeding Series*, pages 89–96. ACM.

Jun Chen, Deyao Zhu, Xiaoqian Shen, Xiang Li, Zechun Liu, Pengchuan Zhang, Raghuraman Krishnamoorthi, Vikas Chandra, Yunyang Xiong, and Mohamed Elhoseiny. 2023a. Minigpt-v2: large language model as a unified interface for vision-language multi-task learning. *CoRR*, abs/2310.09478.

Xi Chen, Xiao Wang, Soravit Changpinyo, A. J. Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, Alexander Kolesnikov, Joan Puigcerver, Nan Ding, Keran Rong, Hassan Akbari, Gaurav Mishra, Linting Xue, Ashish V. Thapliyal, James Bradbury, and Weicheng Kuo. 2023b. PaLI: A jointly-scaled multilingual language-image model. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Feng Gao, Qing Ping, Govind Thattai, Aishwarya N. Reganti, Ying Nian Wu, and Prem Natarajan. 2022. Transform-retrieve-generate: Natural language-centric outside-knowledge visual question answering. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 5057–5067. IEEE.

François Gardères, Maryam Ziaeefard, Baptiste Abeloos, and Freddy Lecue. 2020. ConceptBert: Concept-aware representation for visual question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 489–498, Online. Association for Computational Linguistics.

Michael Glass, Gaetano Rossiello, Md Faisal Mahbub Chowdhury, Ankita Naik, Pengshan Cai, and Alfio Gliozzo. 2022. Re2G: Retrieve, rerank, generate. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2701–2715, Seattle, United States. Association for Computational Linguistics.

Liangke Gui, Borui Wang, Qiuyuan Huang, Alexander Hauptmann, Yonatan Bisk, and Jianfeng Gao. 2022. KAT: A knowledge augmented transformer for vision-and-language. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 956–968, Seattle, United States. Association for Computational Linguistics.

Sebastian Hofstätter, Jiecao Chen, Karthik Raman, and Hamed Zamani. 2023. Fid-light: Efficient and effective retrieval-augmented text generation. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2023, Taipei, Taiwan, July 23-27, 2023*, pages 1437–1447. ACM.

Gautier Izacard and Edouard Grave. 2021a. Distilling knowledge from reader to retriever for question answering. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Gautier Izacard and Edouard Grave. 2021b. 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*, pages 874–880, Online. Association for Computational Linguistics.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 4904–4916. PMLR.

Amita Kamath, Christopher Clark, Tanmay Gupta, Eric Kolve, Derek Hoiem, and Aniruddha Kembhavi. 2022. Webly supervised concept expansion for general purpose vision models. In *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part XXXVI*, volume 13696 of *Lecture Notes in Computer Science*, pages 662–681. Springer.

Jinhyuk Lee, Seongjun Yun, Hyunjae Kim, Miyoung Ko, and Jaewoo Kang. 2018. Ranking paragraphs for improving answer recall in open-domain question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 565–569, Brussels, Belgium. Association for Computational Linguistics.

Guohao Li, Xin Wang, and Wenwu Zhu. 2020. Boosting visual question answering with context-aware knowledge aggregation. In *MM '20: The 28th ACM International Conference on Multimedia, Virtual Event / Seattle, WA, USA, October 12-16, 2020*, pages 1227–1235. ACM.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven C. H. Hoi. 2023. BLIP-2: bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 19730–19742. PMLR.

Yuanze Lin, Yujia Xie, Dongdong Chen, Yichong Xu, Chenguang Zhu, and Lu Yuan. 2022. REVIVE: regional visual representation matters in knowledge-based visual question answering. In *NeurIPS*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023. Improved baselines with visual instruction tuning. *CoRR*, abs/2310.03744.

Tie-Yan Liu. 2009. Learning to rank for information retrieval. *Found. Trends Inf. Retr.*, 3(3):225–331.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. ViLBERT: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 13–23.

Man Luo, Yankai Zeng, Pratyay Banerjee, and Chitta Baral. 2021. Weakly-supervised visual-retriever-reader for knowledge-based question answering. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6417–6431, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yuning Mao, Pengcheng He, Xiaodong Liu, Yelong Shen, Jianfeng Gao, Jiawei Han, and Weizhu Chen. 2021. Reader-guided passage reranking for open-domain question answering. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 344–350, Online. Association for Computational Linguistics.

Kenneth Marino, Xinlei Chen, Devi Parikh, Abhinav Gupta, and Marcus Rohrbach. 2021. KRISP: integrating implicit and symbolic knowledge for open-domain knowledge-based VQA. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021, virtual, June 19-25, 2021*, pages 14111–14121. Computer Vision Foundation / IEEE.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. OK-VQA: A visual question answering benchmark requiring external knowledge. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019, Long Beach, CA, USA, June 16-20, 2019*, pages 3195–3204. Computer Vision Foundation / IEEE.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Adam Roberts, Hyung Won Chung, Anselm Levskaya, Gaurav Mishra, James Bradbury, Daniel Andor, Sharan Narang, Brian Lester, Colin Gaffney, Afroz Mohiuddin, Curtis Hawthorne, Aitor Lewkowycz, Alex Salcianu, Marc van Zee, Jacob Austin, Sebastian Goodman, Livio Baldini Soares, Haitang Hu, Sasha Tsvyashchenko, Aakanksha Chowdhery, Jasmijn Bastings, Jannis Bulian, Xavier Garcia, Jianmo Ni, Andrew Chen, Kathleen Kenealy, Jonathan H. Clark, Stephan Lee, Dan Garrette, James Lee-Thorp, Colin Raffel, Noam Shazeer, Marvin Ritter, Maarten Bosma, Alexandre Passos, Jeremy Maitin-Shepard, Noah Fiedel, Mark Omernick, Brennan Saeta, Ryan Sepassi, Alexander Spiridonov, Joshua Newlan, and Andrea Gesmundo. 2022. Scaling up models and data with `t5x` and `seqio`. *arXiv preprint arXiv:2203.17189*.

Alireza Salemi, Juan Altmayer Pizzorno, and Hamed Zamani. 2023. A symmetric dual encoding dense retrieval framework for knowledge-intensive visual question answering. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2023, Taipei, Taiwan, July 23-27, 2023*, pages 110–120. ACM.

Dustin Schwenk, Apoorv Khandelwal, Christopher Clark, Kenneth Marino, and Roozbeh Mottaghi. 2022. A-OKVQA: A benchmark for visual question answering using world knowledge. In *Computer Vision - ECCV 2022 - 17th European Conference, Tel Aviv, Israel, October 23-27, 2022, Proceedings, Part VIII*, volume 13668 of *Lecture Notes in Computer Science*, pages 146–162. Springer.

Zhenwei Shao, Zhou Yu, Meng Wang, and Jun Yu. 2023. Prompting large language models with answer heuristics for knowledge-based visual question answering. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 14974–14983. IEEE.

Qingyi Si, Zheng Lin, Ming yu Zheng, Peng Fu, and Weiping Wang. 2021. Check it again:progressive visual question answering via visual entailment. 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)*, pages 4101–4110, Online. Association for Computational Linguistics.

Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. 2021. Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, page 2443–2449, New York, NY, USA. Association for Computing Machinery.

Hao Tan and Mohit Bansal. 2019. LXMERT: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5100–5111, Hong Kong, China. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

Shuohang Wang, Mo Yu, Xiaoxiao Guo, Zhiguo Wang, Tim Klinger, Wei Zhang, Shiyu Chang, Gerry Tesauro, Bowen Zhou, and Jing Jiang. 2018. R^3: Reinforced ranker-reader for open-domain question answering. In *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 5981–5988. AAAI Press.

Jialin Wu, Jiasen Lu, Ashish Sabharwal, and Roozbeh Mottaghi. 2022. Multi-modal answer validation for knowledge-based VQA. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pages 2712–2721. AAAI Press.

Zhengyuan Yang, Zhe Gan, Jianfeng Wang, Xiaowei Hu, Yumao Lu, Zicheng Liu, and Lijuan Wang. 2022. An empirical study of GPT-3 for few-shot knowledge-based VQA. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pages 3081–3089. AAAI Press.

Zihao Zhu, Jing Yu, Yujing Wang, Yajing Sun, Yue Hu, and Qi Wu. 2020. Mucko: Multi-layer cross-modal knowledge reasoning for fact-based visual question answering. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence, IJCAI 2020*, pages 1097–1103. ijcai.org.

Honglei Zhuang, Zhen Qin, Rolf Jagerman, Kai Hui, Ji Ma, Jing Lu, Jianmo Ni, Xuanhui Wang, and Michael Bendersky. 2023. RankT5: Fine-tuning T5 for text ranking with ranking losses. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2023, Taipei, Taiwan, July 23-27, 2023*, pages 2308–2313. ACM.

A Experiment Setup

We initialize image-text retrieval module with pretrained ALIGN checkpoint, and we initialize both answer generation and multi-modal reranking module with pretrained PaLI-3b checkpoint.

In the retrieval module, we crop a question image into a series of patches with kernel size 224 with stride 64. We use each image patch to retrieve top-20 candidates and then aggregate candidates from one question image. If there are candidates that are retrieved by multiple image patches in the same image, we will keep the one with highest relevance score. We use aggregated top-20 candidates as candidates set for answer generation training and testing.

For OK-VQA, the multi-modal reranker takes 8500 of examples from training set for training, and the rest of them for model development. For each question, the reranker takes aggregated candidates from top-20 image patch retrieval as the candidate set. At each training step, we will sample 20 candidates for each question and perform pairwise logistics training. We select the reranker checkpoint based on Hits@k. The reranker is then applied to the aggregated image retrieval results to obtain the reranked relevance scores.

Answer generation is trained with batch size as 32 for 10K. Reranker is trained with batch size as 8 for 20K steps. The learning rate is 1e-4. We implement the models based on T5X (Roberts et al., 2022).

B Additional Comparison with Other Ranking Strategies

Ranking Methods	VQA Acc.
Distillation (Izacard and Grave, 2021a)	51.5
RankT5 (Zhuang et al., 2023)	52.3
Reranking	**52.6**

Table 5: Effects of multimodal ranking. We can find that learning reranker using distillation from answer generator can instead hurt the performance. Our multimodal reranker trained with small data provides competitive performance even compare to RankT5 which is pretrained on large amount of data.

We also compare our model to the same multi-modal reranking model architecture trained with knowledge distillation from answer generation (Izacard and Grave, 2021a) and RankT5 (Zhuang et al., 2023) in Table 5. We can find that supervision from knowledge distillation can not provide reliable labels to train a reasonable reranking module. While both text-based reranking and multi-modal reranking can contribute to the performance, and multi-modal reranking can provide better performance. Especially, compared to RankT5 which is pretrained with over 500K items, our reranker only trained with around 8000 items. But it can still achieve competitive performance.

Methods	VQA Acc.
BLIP-2 (Li et al., 2023)	45.9
Flamingo-80b (Alayrac et al., 2022)	50.6
Our model	**52.6**

Table 6: Comparison between multi-modal large models on OK-VQA datasets. We can find that our model provides promising performance compared to the zero-shot performance of those multi-modal large models.

C Comparison With Zero-Shot Multi-Modal Large Models

We also provide additional comparison in Table 6 between some multi-modal large models on OK-VQA, including Flamingo-80b (Alayrac et al., 2022) and BLIP-2 (Li et al., 2023). We report their zero-shot performance compared to our model. The results show that smaller model can still achieve competitive performance when comparing to the zero-shot capability of those large models. We also note that there are some other multi-modal large models such as LLAVA 1.5 (Liu et al., 2023), MiniGPT4-V2 (Chen et al., 2023a), which are instruction tuned with OK-VQA and therefore cannot be directly compared. But in general, our proposed framework can be extended to other multi-modal language models that take the combination of image and text input.