PreFLMR: Scaling Up Fine-Grained Late-Interaction Multi-modal Retrievers

Weizhe Lin, Jingbiao Mei, Jinghong Chen, Bill Byrne
Department of Engineering
University of Cambridge
Cambridge, United Kingdom CB2 1PZ
{wl356, jm2245, jc2124, wjb31}@cam.ac.uk

Abstract

Large Multimodal Models (LMMs) excel in natural language and visual understanding but are challenged by exacting tasks such as Knowledge-based Visual Question Answering (KB-VQA) which involve the retrieval of relevant information from document collections to use in shaping answers to questions. We present an extensive training and evaluation framework, M2KR, for KB-VQA. M2KR contains a collection of vision and language tasks which we have incorporated into a single suite of benchmark tasks for training and evaluating general-purpose multi-modal retrievers. We use M2KR to develop PreFLMR, a pre-trained version of the recently developed Fine-grained Late-interaction Multi-modal Retriever (FLMR) approach to KB-VQA, and we report new state-of-the-art results across a range of tasks. We also present investigations into the scaling behaviors of PreFLMR intended to be useful in future developments in general-purpose multi-modal retrievers. The code, demo, dataset, and pre-trained checkpoints are available at https://preflmr.github.io/.

1 Introduction

Knowledge-based Visual Question Answering (KB-VQA) systems generate answers to queries consisting of questions about given images. Correctly answering these questions requires accessing relevant world knowledge as well as vision and language understanding. Despite their demonstrated abilities in vision and language, recent Large Multimodal Models (LMMs) (Chen et al., 2023a; Driess et al., 2023; Liu et al., 2023a; Zhu et al., 2023; OpenAI, 2023) have performed poorly in recent challenging KB-VQA tasks (Chen et al., 2023b; Mensink et al., 2023a). One promising approach to improve their KB-VQA performance is Retrieval-Augmented Generation (RAG), in which answer generation by LMMs is grounded in relevant documents retrieved from a knowledge base.

The best-performing document retrieval approach for KB-VQA to date is Fine-grained Late-interaction Multi-modal Retrieval (FLMR) (Lin et al., 2023b). FLMR uses multi-dimensional embedding matrices to represent documents and queries and then efficiently computes their relevance scores via late-interaction (Khattab and Zakharia, 2020), thus capturing fine-grained relevance at the token level rather than at the passage level, as in Dense Passage Retrieval (DPR) (Karpukhin et al., 2020). As a late-interaction retriever, FLMR substantially outperforms DPR on a range of KB-VQA tasks, with only minor speed penalties. In all of these methods, model and data size are important considerations. There has been much work in scaling up Large Language Models (LLMs) (Kaplan et al., 2020; Alabdulmohsin et al., 2022; Chen et al., 2023a) and text-based retrieval (Ni et al., 2022), but the scaling properties of these vision and language retrieval systems have not been studied. We therefore investigate the following three aspects of FLMR in KB-VQA.

(1) Vision & Text Encoding: We investigate how KB-VQA performance is affected by scaling the size and complexity of vision and text encoders.

(2) Pre-training: As originally formulated, FLMR employs simple, lightly trained Multi-Layer Perceptrons (MLP). We investigate whether gains can be had through more extensive model pre-training.

(3) Task Diversity: We gather nine open-source vision-language datasets into a suite of benchmark tasks, M2KR, for assessing Multi-task Multi-modal Knowledge Retrieval. M2KR encompasses Image-to-Text, Question-to-Text, and Image&Question-to-Text retrieval tasks, and also includes prompting instructions that can be provided to an LLM for each of the component tasks. General purpose multi-modal retrieval models can be created by...
training on the entirety of the M2KR training data and these models can then be evaluated on any or all of the included tasks. Models can further be fine-tuned for specific M2KR tasks using the task-specific tuning data included in the collection.

We show that M2KR can be used in training an FLMR-based RAG LLM for multi-task multi-modal retrieval. We refer to this model as Pre-FLRM (for Pre-trained FLMR). PreFLMR can be used directly in its pre-trained form for multi-task multi-modal retrieval. PreFLMR can also be fine-tuned for specific task-specific performance. In both uses we find that PreFLMR gives us substantial gains across the M2KR tasks.

Contributions of this paper are:

- The M2KR task suite encompassing nine datasets and three types of retrieval tasks for training and evaluating general-purpose vision-language retrievers. We create M2KR by repurposing various vision and language data sets that might not be originally created for knowledge-based visual question answering, thus ensuring a rich and diverse collection.
- PreFLMR, a strong multi-modal retriever pre-trained on a vision-language corpus of over ten million items. We show that PreFLMR performs well across a range of knowledge retrieval tasks when given the appropriate instructions. We will release PreFLMR upon publication.
- A study of the scaling behaviour of FLMR in terms of its model parameters and training data. To our knowledge, this is the first systematic study of scaling in late-interaction based vision-language retrievers and should provide empirical guidance for future work.

## 2 Related Work

### Document Retrieval

DPR has become a cornerstone in knowledge-intensive tasks (Chen et al., 2017; Izacard and Grave, 2021; Guu et al., 2020; Lee et al., 2019; Lewis et al., 2020) as well as in KB-VQA tasks due to its fast and precise retrieval capabilities (Karpukhin et al., 2020; Gui et al., 2021; Luo et al., 2021; Lin and Byrne, 2022; Wu and Mooney, 2022). Recent developments in retrieval methods, particularly Late Interaction models (Khattab and Zaharia, 2020; Santhanam et al., 2022b), have shown notable performance gains over DPR, albeit with some efficiency trade-offs (Lin et al., 2023a,b). In multi-modal retrieval, FILIP (Yao et al., 2022) used pre-trained late interaction models for single-modal image-text retrieval, while FLMR (Lin et al., 2023b) extended the approach to multi-modal retrieval for KB-VQA with finer-grained visual and text features. This paper further extends FLMR and explores its scaling properties in multi-modal retrieval. Similar to our M2KR benchmark, a concurrent work (Wei et al., 2023) introduces M-Beir, which combines several retrieval tasks and can be used to train and evaluate universal multi-modal retrievers.

Another line of relevant research is KB-VQA retrieval involving Named Entities, where retrieved documents must identify the person in the image. For example, on ViQuAE (Lerner et al., 2022), Lerner et al. (2023) trains the retriever with a multimodal inverse cloze task, while Lerner et al. (2024) shows that combining mono- and cross-modal retrieval improves performance. Both use a weighted sum of BERT (Devlin et al., 2019) and CLIP (Radford et al., 2021) embeddings, while our work trains a single multi-modal late-interaction retriever.

### Knowledge-based VQA Systems

Recent multi-modal systems have significantly improved in complex tasks like OKVQA (Schwenk et al., 2022) that require external knowledge sources (Narasimhan et al., 2018; Garderes et al., 2020; Li et al., 2020; Wu et al., 2022; Marino et al., 2021; Chen et al., 2023d; Gao et al., 2022; Gui et al., 2021; Hu et al., 2023b; Rao et al., 2023). Systems like KAT (Gui et al., 2021) and REVIVE (Lin et al., 2022) used LLMs (e.g. GPT-3) for generating candidate answers. Challenges remain in answering more knowledge-intensive questions (Chen et al., 2023b; Mensink et al., 2023a), underscoring the need for robust document retrieval. Mensink et al. (2023a) showed that even state-of-the-art LLMs perform poorly on difficult KB-VQA questions, with an accuracy of under 20% when retrieval is not incorporated. RA-VQAv2 (Lin et al., 2023b) and prior work (Lin and Byrne, 2022; Luo et al., 2021; Qu et al., 2021; Gao et al., 2022; Hu et al., 2023b; Mensink et al., 2023a) demonstrated strong performance in KB-VQA by using external knowledge databases.

### Scaling Retrieval Systems

Previous work has explored scaling laws in language/vision systems (Kaplan et al., 2020; Alabdulmohsin et al., 2022), revealing correlations between model performance, computation, number of parameters, and dataset sizes. In retrieval, Ni et al. (2022) and Hu et al. (2023b) both observe improvements in DPR-like...
models with one-dimensional embeddings by increasing the size of language/vision encoders. This paper reports similar scaling investigations in multimodal late-interaction retrieval.

3 The M2KR Benchmark Suite

Current multi-modal retrievers are typically trained and evaluated on a single dataset only. To properly study general-purpose multi-modal retrievers, we introduce the Multi-task Multi-modal Knowledge Retrieval (M2KR) benchmark suite. We convert nine diverse datasets, originally designed for vision and language tasks such as image recognition, image captioning, and conversational interactions, into a uniform retrieval format. Details of the pre-processing steps, data partition, and prompting instructions are provided in Appendix A, but we note here that re-purposing these datasets into a single consistent collection for knowledge-based visual question answering represents a non-trivial effort. M2KR will be released with our models.

3.1 Tasks and Datasets

Table 1 shows the composition of M2KR. We pre-process the datasets into a uniform format and write several task-specific prompting instructions for each dataset. The M2KR benchmark contains three types of tasks:

**Image to Text (I2T) retrieval.** These tasks evaluate the ability of a retriever to find relevant documents associated with an input image. Component tasks are WIT (Srinivasan et al., 2021), IGLUE-en (Bugliarello et al., 2022), KVQA (Shah et al., 2019), and CC3M (Sharma et al., 2018). CC3M is included in the M2KR training set to improve scene understanding but not in the validation/test set as the task concerns caption generation, not retrieval. The IGLUE test set, which is a subset of WIT and has an established benchmark, is included to enable comparison with the literature. The KVQA task, initially designed as a KB-VQA task, has been re-purposed into an I2T task for our modelling purposes (Appendix A.1.3).

**Question to Text (Q2T) retrieval.** This task is based on MSMARCO (Bajaj et al., 2018) and is included to assess whether multi-modal retrievers retain their ability in text-only retrieval after any retraining for images.

**Image & Question to Text (IQ2T) retrieval.** This is the most challenging task which requires joint understanding of questions and images for accurate retrieval. It consists of these subtasks: OVEN (Hu et al., 2023a), LLaVA (Liu et al., 2023b), OKVQA (Schwenk et al., 2022), Infoseek (Chen et al., 2023c) and E-VQA (Mensink et al., 2023b). We note in particular that we convert LLaVA, a multi-modal conversation dataset, into a multi-modal retrieval task (Appendix A.3.1).

The training/validation/test examples are down-sampled from the respective sets of the original datasets. We take test examples from the original validation sets for LLaVA and Infoseek since LLaVA has no test sets and the test set annotation of Infoseek has not been released. We limit the maximum test samples to 5,120 for each dataset to allow faster performance tests on all 9 datasets. Data preprocessing and partitioning details are in Appendix A. We further verified that there are no identical images between the training and test sets by checking the MD5 of the images, thereby preventing data contamination during training. We use the validation splits to select hyperparameters for the models which can be found in detail in Appendix B.2

3.2 Evaluation

We use Recall@K (R@K), which measures whether at least one of the target documents is in the top-K retrieved entries, to evaluate retrieval performance. Additionally, for the datasets Infoseek, E-VQA, and OKVQA, we mainly employ Pseudo Recall/PRecall@K (PR@K) for evaluation. This metric measures whether at least one of the top K documents includes the target answer.\(^1\)

<table>
<thead>
<tr>
<th>Datasets</th>
<th>#Examples</th>
<th>#Passages</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>I2T Retrieval</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WIT</td>
<td>2.8M</td>
<td>4.1M</td>
</tr>
<tr>
<td>IGLUE</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>KVQA</td>
<td>65K</td>
<td>16.3K</td>
</tr>
<tr>
<td>CC3M</td>
<td>595K</td>
<td>-</td>
</tr>
<tr>
<td><strong>Q2T Retrieval</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MSMARCO</td>
<td>400K</td>
<td>8.8M</td>
</tr>
<tr>
<td><strong>IQ2T Retrieval</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>OVEN</td>
<td>339K</td>
<td>10K</td>
</tr>
<tr>
<td>LLaVA</td>
<td>351K</td>
<td>351K</td>
</tr>
<tr>
<td>OKVQA</td>
<td>9K</td>
<td>110K</td>
</tr>
<tr>
<td>Infoseek</td>
<td>100K</td>
<td>100K</td>
</tr>
<tr>
<td>E-VQA</td>
<td>212K</td>
<td>50K</td>
</tr>
</tbody>
</table>

Table 1: Datasets in M2KR Benchmark Suite.

\(^1\)In practice, PR@K more accurately reflects actual retrieval performance and exhibits a stronger correlation with
We use R@10 for WIT and MSMARCO, and R@1 for LLaVA and IGLUE. Other datasets are evaluated with R@5 or PR@5. As in Table 2, we also report the average rank (A.R.) of each model over all datasets to indicate multi-task retrieval performance relative to other models in comparison; lower is better.

3.3 Baselines

For each dataset, we show the best published results in recent literature as baselines, if available (Table 2). For datasets without previous results such as LLaVA and OVEN, we use our replication of CLIP (Radford et al., 2021) and FLMR as zero-shot baselines following Lin et al. (2022).

4 PreFLMR Architecture and Training

Our architecture generally follows that of FLMR (Lin et al., 2023b) as shown in Fig. 1. PreFLMR uses token embedding matrices Q and D to represent query and document, respectively. Given a query \( q \) consisting of texts \( q \) and an image \( I \), PreFLMR uses a language model \( F_L \) to obtain embeddings of all tokens in \( q \), a vision model \( F_V \) to obtain embeddings of \( I \), and a mapping structure \( F_M \) to project image embeddings into the text embedding space. All token-level embeddings are concatenated to form the query representation \( Q \). The document matrix \( D \) is obtained similarly with the language model \( F_L \) but without visual features.

The relevance score \( r(q, d) \) is computed via late-interaction (Khattab and Zaharia, 2020) between \( Q \) and \( D \), aggregating the maximum dot products over all query tokens with respect to all document tokens (Eq. 9). \( l_Q \) and \( l_D \) denote the total number of tokens in query \( q \) and document \( d \), respectively.

\[
r(q, d) = \sum_{i=1}^{l_Q} \max_{j=1}^{l_D} Q_j^\top D_i
\]  

PreFLMR improves over FLMR in the following aspects: (1) While FLMR only uses the [CLS] embedding from ViT as the image representation, in PreFLMR we additionally extract image embeddings of image patches from ViT’s penultimate layer to obtain a detailed visual representation. (2) We introduce Transformer blocks with cross-attention into the mapping structure to obtain query-aware visual representation. The Transformer blocks take the image patch embeddings as input, and use cross-attention to integrate the features of the text encoder. This allows PreFLMR to attend to different aspects of the image under different queries. These Transformer blocks are placed in parallel with FLMR’s 2-layer MLP mapping structure. (3) We append task-specific instructions to the text query to distinguish between tasks. The list of instructions for each task can be found in Appendix A. For each query, the instruction is randomly sampled from the corresponding instruction list. Instruction tokens are masked in computing relevance score. For Q2T retrieval training, we feed a blank image as PreFLMR’s image input. For I2T retrieval training, we use instructions as text input to PreFLMR.

PreFLMR training and inference follow that of FLMR. When training on data consisting of several datasets, we randomly shuffle the entire training data and only use in-batch negative examples from the same corpus. Post-training, all documents are indexed through PLAID (Santhanam et al., 2022a) for efficient late-interaction retrieval. For detailed evaluation of retrieval efficiency, we refer readers to Lin et al. (2023b).

The detailed formal expression of the entire model can be found in Appendix B.4.

4.1 Training Procedures

PreFLMR’s pre-training involves four stages.

Stage 0: Text Encoder Pre-training. We train ColBERT following Khattab and Zaharia (2020) on the MSMARCO dataset to obtain the initial checkpoint for PreFLMR’s text encoder \( F_L \). This is a straightforward replication of ColBERT used as an initial text encoder as was done in FLMR, but also allowing for size variations.

Stage 1: Training the Mapping Structure. In this stage, we only train the mapping structure \( F_M \), keeping the language and vision models frozen. This approach is an extension of the FLMR methodology, incorporating a larger dataset and an additional cross-attention mapping layer. The training is performed on the IQ2T dataset (LLaVA, OVEN), I2T datasets (WIT, CC3M, KVQA), and Q2T dataset (MSMARCO). Our objective is to encompass all three task types in M2KR without the need to optimize the data mixing ratio or manually select datasets to achieve an effective mapping structure. This strategy is inspired by pre-
previous studies (Lin et al., 2023b; Zhu et al., 2023; Liu et al., 2023b), which utilized relatively simple multi-modal tasks to develop image-to-text mappings.

We mask the late-interaction token embeddings in query matrix Q that are produced by the language model (not the token embeddings at the input embedding layer). This encourages the Transformer cross-attention layer to integrate information from its textual inputs and enables PreFLMR to perform IQ2T, I2T, and Q2T retrieval when provided with the appropriate instructions for each task.

**Stage 2: Intermediate KB-VQA Pre-training.**

We tune the text encoder $F_L$ and the mapping structure $F_M$ on the E-VQA dataset, a large and high quality KB-VQA dataset, to enhance PreFLMR’s retrieval performance. Including an intermediate pre-training stage to align the model with in-domain data has been well-explored in the literature (e.g., Google’s TAPAS (Eisenschlos et al., 2020)). We opt for a straightforward procedure to train on E-VQA in the intermediate stage because of its diversity, increased difficulty, and larger quantity compared to other KB-VQA datasets. Specifically, E-VQA requires recognition of less common entities such as spotted hyenas and relies on more specialized domain knowledge such as American landmarks, making it good for retrieval training. This design choice is well-supported by experimental results (Table 2 #8 vs #5, #3 vs #2) and we provide detailed analysis in Sec. 5.6.

**Stage 3: Full-scale Fine-tuning.** We train on the entire M2KR corpora, including OKVQA and Infoseek. This stage is straightforward multi-task learning. We tune the entire model except the vision encoder $F_V$. We adjust the dataset proportions to ensure balanced learning on these datasets of varying sizes (Appendix B.1). Additionally, we use separate text encoders to encode queries and documents; their parameters were shared in previous steps.

### 4.2 Training Configurations

We use the Adam optimizer (Kingma and Ba, 2015) with a fixed learning rate of $10^{-4}$ for the mapping structures and $10^{-5}$ for other parameters in all experiments. Training was run up to 300k, 220k, 12k, and 50k steps in the four stages, respectively. Full training configurations (including the hyperparameters for downstream VQA fine-tuning) can be found in Appendix B.2.

### 5 Experiments and Results

In this section we present results of scaling PreFLMR components (Sec. 5.2, 5.4), analyze the effect of each training stage (Sec. 5.3, 5.6), and evaluate on the downstream KB-VQA tasks (Sec. 5.5). We summarize our findings in Sec. 5.7. Multi-task performance refers to PreFLMR results, i.e. Stages 0, 1, 2, and 3, without any single-task fine-tuning.
5.1 Model Variants

We experiment with a range of model configurations. Model sizes range from BERT-Small (28.8M), BERT-Medium (41.1M), BERT-Base (110M) to BERT-large (340M). ColBERT text encoders are denoted as "[BERT size]-[pre-training scheme]". There are two ColBERT pre-training schemes: "v1" (Khattab and Zaharia, 2020) and "v2" (Santhanam et al., 2022b). "v2" yields a better performing model than "v1" as evaluated on MSMARCO. We compare models initialized from "v1" and "v2" checkpoints to investigate how the performance of the initial uni-modal text retriever affects the final multi-modal vision-language retriever. Except for "Base-v2", all ColBERT variants are trained using our replication of ColBERT following the "v1" pre-training scheme.2

2The training code of "v2" has not been released officially.

5.2 PreFLMR Performance

The best-performing PreFLMR model (ViT-G + Base-v2) outperforms other variants on most of M2KR benchmark (Table 2, #13). Without single-task fine-tuning, PreFLMR outperforms baseline models optimized for the individual tasks on 7 out of 9 datasets, showcasing its capability as a general visual-language retriever. We now analyze how each PreFLMR component affects performance.

Vision Encoder Scaling. Scaling ViT from ViT-B (88M) to ViT-G (1.8B) while keeping the text encoder fixed brings about substantial performance gain across all tasks (Table 2 #2, #5, #12, #13), e.g. 48.8 to 59.6 on Infoseek and 67.9 to 73.1 on E-VQA. The gain is greater when upgrading ViT-B to ViT-L with recall improvements of ∼10% on WIT, KVQA, OVEN, and Infoseek, showing the benefit of using better vision encoders. In addition, Fig. 3 in the appendix illustrates performance gains in scaling the vision encoder with a radar plot. However, the performance plateaus when scaling ViT to H and G. This observation aligns with results reported in the literature. OpenCLIP (Cherti et al., 2023) and BLIP2 (Li et al., 2023) have reported marginal or no performance improvement when scaling beyond ViT-L across several datasets. A plausible explanation is that the ViT model is not pre-trained on domain-specific data, it may struggle to make fine distinctions.

Text Encoder Scaling. Scaling up the text encoder from BERT-Small-v1 to Medium-v1 to Base-v1 (Table 2 #9, #10, #4) yields substantial performance gain (A.R. 8.3, 8.2, and 5.6). However, we find that further scaling to Large-v1 (#11) adversely impacts the performance (A.R. decreased to 6.6). We attribute this to overfitting and unstable training for large models given the available data (Appendix B.3). The results suggest that BERT-Base (110M) is adequate for building a capable vision-language retriever.

Improving Text Encoder. Compared to PreFLMR models initialized from Base-v1, models initialized from Base-v2 have better multi-tasking performance indicated by better A.R. (Table 2 #1 vs #2 and #4 vs #5). The gain from improving the text encoder is more substantial when using the “ViT-L” vision model (-2.4 A.R.) compared to using “ViT-B” (-0.8 A.R.), indicating that the text encoder is relatively weak as the vision model improves.

5.3 Performance of Each PreFLMR Stage

In this section, we analyze intermediate performance in the earlier stages of pre-training to better understand the scaling behaviour of PreFLMR.

Text Encoder Pre-training. We train “ColBERT-v1” at different sizes and evaluate on the MSMARCO dataset. Table 3 shows larger model sizes consistently yield better text retrieval performance. In contrast to the multi-modal case, scaling up to “Large-v1” does not destabilize training and leads to better performance compared to “Base-v1”.

Training the Mapping Structures. Table 4 details system performance after Stage 1 training, in which only the vision-language mapping structure is trained. Similar to Sec.5.2, scaling up the vision encoder improves performance across tasks. PreFLMR exhibits strong zero-shot KB-VQA performance at this preliminary stage (50.87 in Infoseek, 42.44 in E-VQA, and 52.14 in OKVQA). After Stage 1, PreFLMR with ViT-G performs worse than other variants on IGLUE, E-VQA and OKVQA. However, it attains the best performance on these datasets after Stage 3. This suggests that tuning the mapping structure alone is not enough to fully utilize larger vision models.

Intermediate Pre-training. Stage 1 improves performance on KB-VQA tasks (Table 2 #3 vs #2
Table 2: PreFLMR performance on all datasets. PR stands for Pseudo Recall. Best multi-task performance is in bold and best downstream fine-tuning performance is underlined. For the vision encoder, we compare ViT-B (B), ViT-L (L), ViT-H (H) and ViT-G (G). For the text encoder, we compare Base-v1 (B-v1), Base-v2 (B-v2), Small-v1 (S-v1), Medium-v1 (M-v1), and Large-v1 (L-v1). A.R.: Average Rank against all other models on all tasks. For baselines, we show: GIVL (Yin et al., 2023) for IGLUE; ColBERTv2 for MSMARCO (MM); FLMR (Lin et al., 2023b) for Infoseek and OKVQA; and Google Lens (Google) for E-VQA. We follow the procedure as detailed in the Appendix C of the E-VQA paper (Mensink et al., 2023b) to use CLIP as a zero-shot retriever.

5.4 Ablation Studies

Instructions. Removing instructions (Table 2 #7) results in much worse overall performance, with the WIT recall rate reduced to 13.3. This shows that instructions are necessary for multi-task learning and that our instruction scheme works well (the full list of instructions is given in Appendix A).

Pre-training Datasets. As shown in Table 5, adding CC3M to training improves performance on all metrics, showing that learning to understand scene via captioning datasets is beneficial. Removing either LLaVA or MSMARCO harms zero-shot KB-VQA performance (~3.0 in Infoseek), noting that Infoseek is not used in this stage. Training on these datasets facilitates learning question-aware visual representations as the cross-attention in the mapping structure must attend to the text input to perform well on these tasks. Omitting knowledge-intensive datasets (OVEN and KVQA) negatively impacts the zero-shot performance on Infoseek, showing the importance of in-domain data.

Mapping Structure Scaling. Table 6 illustrates the impact of scaling up the mapping structure under two PreFLMR configurations. Increasing cross-attention layers from 1 to 4 marginally improves LLaVA performance (+0.5, approx.), but adversely

Table 5: Ablation study on Stage 1 pre-training datasets. The model is ViT-B + Base-v1. We evaluate systems on Infoseek in zero-shot mode though it is not used in Stage 1 training.

and #8 vs #5). Although we are only training on E-VQA, the score on other KB-VQA tasks (Infoseek, KVQA, OKVQA) increases by ~1% or more. This shows that E-VQA is an appropriate corpus for training a general-purpose knowledge retriever. We analyze the gain from intermediate pre-training in more detail in Sec.5.6.
Table 3: Text encoder pre-training results evaluated on the full MS-MARCO test set.

<table>
<thead>
<tr>
<th>Model</th>
<th>MRR@10</th>
<th>Recall@50</th>
</tr>
</thead>
<tbody>
<tr>
<td>Small-v1 (28.8M)</td>
<td>34.5</td>
<td>79.8</td>
</tr>
<tr>
<td>Medium-v1 (41.4M)</td>
<td>35.5</td>
<td>81.4</td>
</tr>
<tr>
<td>Base-v1 (110M)</td>
<td>35.8</td>
<td>82.4</td>
</tr>
<tr>
<td>Large-v1 (345M)</td>
<td>37.0</td>
<td>83.2</td>
</tr>
<tr>
<td>Base-v1 (official)</td>
<td>36.0</td>
<td>82.9</td>
</tr>
<tr>
<td>Base-v2 (official)</td>
<td>39.7</td>
<td>86.8</td>
</tr>
</tbody>
</table>

Table 4: PreFLMR performance after Stage 1. Infoseek, E-VQA, and OKVQA are tested in zero-shot mode. A.R.: Average Rank against all other models on all tasks. LLa.- LLaVA; Info.- Infoseek; OK. - OKVQA.

<table>
<thead>
<tr>
<th>Model</th>
<th>NTR</th>
<th>WIT</th>
<th>LLaVA</th>
<th>Infoseek</th>
<th>E-VQA</th>
</tr>
</thead>
<tbody>
<tr>
<td>ViT-B + Base-v1</td>
<td>1L</td>
<td>34.1</td>
<td>50.8</td>
<td>42.7</td>
<td></td>
</tr>
<tr>
<td>ViT-B + Base-v1</td>
<td>4L</td>
<td>29.0</td>
<td>51.4</td>
<td>40.8</td>
<td></td>
</tr>
<tr>
<td>ViT-L + Base-v2</td>
<td>1L</td>
<td>49.6</td>
<td>51.2</td>
<td>48.7</td>
<td></td>
</tr>
<tr>
<td>ViT-L + Base-v2</td>
<td>4L</td>
<td>45.9</td>
<td>51.7</td>
<td>46.8</td>
<td></td>
</tr>
</tbody>
</table>

5.5 Retrieval Augmented Visual Question Answering with PreFLMR

Table 6: Performance of adding more Transformer layers to the mapping structure. NTR is the number of Transformer layers in the mapping structure.

<table>
<thead>
<tr>
<th>Model</th>
<th>OKVQA</th>
<th>Infoseek</th>
<th>E-VQA</th>
</tr>
</thead>
<tbody>
<tr>
<td>Baseline</td>
<td>66.10</td>
<td>21.80</td>
<td>48.80</td>
</tr>
<tr>
<td>Baseline model</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ViT-B + Base-v1</td>
<td>60.20</td>
<td>50.70</td>
<td>45.40</td>
</tr>
<tr>
<td>ViT-L + Base-v2</td>
<td>61.88</td>
<td>30.65</td>
<td>54.45</td>
</tr>
<tr>
<td>RA-VQA v2 w/ FLMR</td>
<td>60.75</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>RA-VQA v2 w/ PreFLMR</td>
<td>55.44</td>
<td>21.78</td>
<td>19.80</td>
</tr>
</tbody>
</table>

Table 7: Downstream KB-VQA performance when RA-VQA v2 (Lin et al., 2023b) is equipped with PreFLMR and fine-tuned on the target M2KR’s KB-VQA sub-tasks. AVIS (Hu et al., 2024) is a recently published hybrid system that leverages many planning stages to solve KB-VQA questions, which we include for reference. Performance on Infoseek and E-VQA may not be directly comparable to results in the literature.5

We build on RA-VQA v2 (Lin et al., 2023b), a strong retrieval-augmented visual question answering system to tackle OKVQA, Infoseek, and E-VQA. We fine-tune the best-performing PreFLMR variant on the target retrieval task (ViT-G + Base-v2, Table 2 #14) and follow RA-VQA v2 to fine-tune a BLIP-2 answer generator on the target M2KR KB-VQA task.5 Following previous literature (Schwenk et al., 2022; Chen et al., 2023c; Mensink et al., 2023b), we use VQA score, Accuracy, and BERT matching (BEM) (Bulian et al., 2022) to evaluate performance on OKVQA, Infoseek, and E-VQA, respectively. A brief summary of the systems shown in Table 7: PaLM-E (Driess et al., 2023), PALI-X (Chen et al., 2022) and PaLM-B (Anil et al., 2023) are large multi-modal models with 562B, 55B, and 1T parameters, respectively. The E-VQA SOTA (Mensink et al., 2023b) uses Lens (Google), the Google API for image retrieval. AVIS (Hu et al., 2024) is a hybrid system with many components (such as PaLI, PaLM, and Google Lens&Web Search API) and planning stages powered by LLMs. We note that PreFLMR could be used as part of the AVIS pipeline to enhance its ability to fetch relevant documents given questions and images. As shown in Table 7, compared to models without retrieval, PreFLMR improves performance by approximately 6% on OKVQA, 9% on Infoseek, and 34% on E-VQA. These results highlight the effectiveness of PreFLMR in document retrieval for KB-VQA tasks.

On OKVQA, the performances of RA-VQA v2 (PreFLMR) and RA-VQA v2 (FLMR) are similar. Table 2 #13 shows that PreFLMR attains similar Recall@5 as FLMR on OKVQA even though it has a much larger vision encoder. As a possible explanation, compared to E-VQA and Infoseek, the knowledge required to answer OKVQA question is less specialized and many OKVQA questions can be answered without document retrieval (Mensink et al., 2023b). See Appendix E for qualitative analysis.5

5We note that this work was conducted during the early stage of the release of Infoseek and E-VQA. We prepared the data splits according to the need for retrieval training following Appendix A. The systems are trained and evaluated on the data splits provided in M2KR to show the improvement relative to systems without retrieval.
Another possibility is that, compared to E-VQA and Infoseek where the ground-truth document is provided for each question, the OKVQA training set does not provide ground-truth knowledge documents. The retriever uses pseudo-relevant documents in training that contain the target answer but these may not be truly useful for answering the question. This is evidence that data quality should be improved along with model scaling.

5.6 Analysis of Intermediate Pre-training

Sec. 5.3 shows that Stage 2 Intermediate Pre-training improves the performance as evaluated by task-specific metrics. In this section, we further quantify the gains from Stage 2 for each dataset and more clearly show that KB-VQA tasks benefit more from Stage 2 than other tasks. We use the difference in minimal validation loss\(^6\) achieved on each dataset starting from checkpoints before or after Stage 2 Intermediate Pre-training as a measure of benefit. This enables comparison of tasks with different performance metrics. Intuitively, a larger absolute difference in validation loss indicates that the dataset benefits more from the Intermediate Pre-training stage.

Figure 2: Change in Stage 3 validation loss when initialized from Stage 2 checkpoints after \(N_{\text{inter}}\) steps of intermediate pre-training. A large difference indicates a greater gain from intermediate pre-training.

Figure 2 plots the difference in validation loss of every dataset when the starting checkpoints have undergone \(N_{\text{inter}}\) intermediate pre-training steps using either BERT-medium or BERT-base as the text encoder backbone. As expected, starting from E-VQA-pre-trained checkpoints yields lower validation loss in knowledge-intensive tasks such as OKVQA, KVQA, and OVEN after the same number (5,000) of fine-tuning steps. Performance on these datasets indeed sees more gain from Stage 2 training (Table 2, #5 v.s. #8). Figure 2 also indicates the existence of an optimal \(N_{\text{inter}}\), beyond which the model overfits to E-VQA, harming performance on other datasets. The larger PreFLMR model with BERT-base text encoder overfits faster than PreFLMR with BERT-medium (\(N_{\text{inter}} \approx 15,000\) versus \(N_{\text{inter}} \approx 10,000\)). We use V-Entropy (Xu et al., 2020) to formalize our analysis as an empirical measure of mutual information between datasets in Appendix D.

5.7 Summary of Findings

We summarise the results of our investigations into scaling behaviour as follows:

- The text encoder size need not exceed that of BERT-base (110M) to achieve competitive multi-modal retrieval performance (Sec.5.2).
- Scaling up the vision encoder from ViT-B to ViT-G yields substantial gains (Sec.5.2).
- Scaling up the mapping structure does not improve performance (Sec.5.4).
- Intermediate pre-training on high-quality in-domain data (E-VQA) effectively improves retrieval performance across KB-VQA tasks (Sec.5.3, 5.6).
- Strong knowledge retrievers boost performance on challenging KB-VQA tasks such as OKVQA, Infoseek, and E-VQA via Retrieval-Augmented Generation (Sec.5.5).
- Ground-truth document labels are important to make full use of large models in training multi-modal retrievers (Sec.5.5).

6 Conclusion

This work has studied the scaling behaviour of state of the art multi-modal document retrieval systems, with a focus on enhancing fine-grained late-interaction retrieval for knowledge-based visual question answering. We contribute a comprehensive training and evaluation framework, M2KR, for general-purpose multi-modal knowledge retrieval. The PreFLMR system we train in the M2KR framework yields excellent retrieval performance across a range of tasks and can also serve as a base for further task-specific fine-tuning.
Limitations

Limited by available computational resources, we leave several further investigations as future work:
(1) The CLIP-ViT models (Cherti et al., 2023) were not pre-trained on in-domain data of knowledge-intensive tasks. Further training may enhance the model’s ability to recognize a broader range of objects; (2) Advanced training approaches beyond contrastive learning, such as score distillation (Santhanam et al., 2022b), could be explored to further enhance retrieval performance; (3) Investigating a more optimal mix proportion of datasets with varying sizes also warrants further exploration.

Ethics Statement

Our proposed model retrieves documents without generating new content. We acknowledge the potential for the retrieved documents to include inappropriate information if the document database lacks adequate filtering. Consequently, extra care must be taken to ensure the sanitization of the document database, particularly when employing this model in applications involving direct interaction with real users.

Acknowledgments

This work was supported in part by the AWS Cloud Credit for Research programme.

Weizhe Lin is supported by a Research Studentship funded by Toyota Motor Europe (RG92562(24020)) for the undertaking of the PhD in Engineering at the University of Cambridge.

Jingbiao Mei is supported by Cambridge Commonwealth, European and International Trust for the undertaking of the PhD in Engineering at the University of Cambridge.

Jinghong Chen is supported by the Warwick Postgraduate Studentship from Christ’s College and the Huawei Hisilicon Studentship for the undertaking of the PhD in Engineering at the University of Cambridge.

Prof. Bill Byrne holds concurrent appointments as a Professor of Information Engineering at Cambridge University and as an Amazon Scholar. This publication describes work performed at Cambridge University and is not associated with Amazon.

We would also like to thank all the reviewers for their knowledgeable reviews.

References


Emanuele Bugliarello, Fangyu Liu, Jonas Pfeiffer, Siva Reddy, Desmond Elliott, Edoardo Maria Ponti, and Ivan Vulić. 2022. Igglue: A benchmark for transfer


Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 2013. 3d object representations for fine-grained categorization. In 4th International IEEE Workshop on 3D Representation and Recognition (3dRR-13), Sydney, Australia.


A Datasets details

This section outlines the preprocessing methods used to convert various datasets into formats suitable for retrieval tasks. Table 8 provides examples from each dataset, demonstrating the transformation from their original to the adapted structure. Subsequent subsections detail the specific preprocessing steps for each dataset. The M2KR dataset is available at Huggingface Hub.

A.1 I2T Retrieval

A.1.1 WIT

WIT (Srinivasan et al., 2021) is a corpus based on Wikipedia with image-text pairs, where the text is the Wikipedia passage associated with the image. To enhance data quality, we exclusively select image-text pairs where the images are the main/title images of their respective Wikipedia documents, and we limit our scope to English-language documents.

Our training set, comprising 2.8 million examples, is sourced from the original WIT training set. 20,102 and 5,120 examples from the original WIT validation set are selected to build the validation set and test set in our M2KR benchmark, respectively. The test corpus includes all documents from the original WIT validation and test sets. This setting ensures that there is no overlap between different sets.

Each image-document pair is paired with a randomly selected instruction from our set of templates. The task is to retrieve the correct document from the test corpus, given the image and instruction.

A.1.2 IGLUE

The IGLUE English retrieval test set (Bugliarello et al., 2022), which is a subset of the WIT test set and has an established benchmark for image-to-text retrieval, is included to enable comparison with models in previous literature. Following Bugliarello et al. (2022), the test set contains 685 unique images and 1,000 Wikipedia passages. The task is similar to WIT: using the image and the instruction to retrieve the corresponding Wikipedia passage.

Instruction templates for WIT and IGLUE:

• <Image> Identify the document that is connected to this image.
• <Image> Provide information about the document linked to this image.
A.1.3 KVQA

KVQA (Shah et al., 2019) is a dataset containing a rich collection of entities representing famous individuals. The KVQA task, initially designed as a KB-VQA task, has been re-purposed into an I2T task for our modelling purposes. This adaptation is based on our findings that using images as queries alone suffices to retrieve the documents containing the correct identities. In our context, where the primary focus is on document retrieval, the original questions are unnecessary. Our reformulated task for KVQA is to retrieve the details of famous people like gender, nationality, birthplace, and employment history based solely on their images. The training set is downsampled from the KVQA original training set by removing repeated examples of the same famous individuals. We transformed the structured entities such as gender and nationality into passages. For example, “nationality: America; date of birth: dd/mm/yyyy; ...” is serialized as “nationality is America, date of birth is dd/mm/yyyy, ...”.

The training corpus is composed of all the documents that appear in the original KVQA training set. For the validation/test set, we selected a subset of 13,365/5,120 samples from the original KVQA validation set. Correspondingly, the test corpus encompasses all documents found in the original KVQA validation set.

The instruction we use for KVQA is: <Image> Provide a brief description of the image and the relevant details of the person in the image.

A.1.4 CC3M

CC3M (Sharma et al., 2018) is a dataset consisting of a vast collection of image-caption pairs. Instead of utilizing the entire dataset comprising 3 million pairs, we adopt the downsampling methodology as delineated in LLaVA’s work (Liu et al., 2023b), resulting in a reduced dataset of approximately 595K.

We reformulate the image-caption pairs into image-to-text retrieval tasks in our pre-training. To construct the training corpus, we treat each caption as an individual document linked to its corresponding image. The task then involves retrieving the most relevant caption for a given image, guided by a set of randomly selected instructions. Since CC3M is originally an image captioning task, we do not validate or test our retriever on CC3M.

Instruction templates for CC3M

- Describe the image concisely.
- Provide a brief description of the given image.
- Offer a succinct explanation of the picture presented.
- Summarize the visual content of the image.
- Give a short and clear explanation of the subsequent image.
- Share a concise interpretation of the image provided.
- Present a compact description of the photo’s key features.
- Relay a brief, clear account of the picture shown.
- Render a clear and concise summary of the photo.
- Write a terse but informative summary of the picture.
- Create a compact narrative representing the image presented.

A.2 Q2T Retrieval

A.2.1 MSMARCO

MSMARCO (Bajaj et al., 2018) stands for Microsoft Machine Reading Comprehension dataset. It is a text-only dataset with around 1 million questions and 8 million passages. At stage 0, we train according to ColBERT-v1 by Khattab and Zaharia (2020). For later stages, we downsample the dataset to 400K questions to balance between the multi-modal tasks and unimodal tasks. For the training corpus, we still use the full 8 million passages. For testing, we select 6,980 and 5,120 samples from the original MSMARCO validation set and sample 400K passages to retrieve from and ensure the subset contains all ground-truth passages.

Instruction templates for MSMARCO:
• Retrieve the document that answers this question. 
• Find the document that is most relevant to the question. 
• Obtain the document that resolves this query. 
• Acquire the document that elucidates this question. 
• Choose the document most relevant to the query. 
• Identify the document most applicable to the question. 
• Extract the document that addresses the query.

A.3 IQ2T Retrieval

A.3.1 LLaV A

The LLaV A instruction following dataset contains GPT-3.5 generated high-quality conversation about an image between a human and an AI assistant. There are around 150K rounds of conversations. We took each conversation (each question from the human and the answer from the AI assistant) as a separate sample. This results in a total of 356K samples. Since there are no original validation or test sets associated with the LLaV A, we manually split the sample pool into 351K training examples and 5,120 test examples.

The task is reformulated to an Image&Question to Text retrieval task. The training corpus and test corpus each contain the associated answers as passages to be retrieved by the image and question pairs. We use two types of instruction templates depending on the preciseness of the question:
• Provide a brief description of the image along with the following question:
• Provide a concise explanation of the image along with the following question:

A.3.2 OVEN

OVEN is a dataset targeting open-domain visual entity recognition. The dataset consists of two splits: entity set and query set. The entity set is derived from image classification datasets such as INaturalist2017 (Cui et al., 2018), Food-101 (Bossard et al., 2014), Cars196 (Krause et al., 2013) and Google Landmarks Dataset v2 (Weyand et al., 2020). The query set is derived from VQA datasets such as VQA v2 (Goyal et al., 2017) and OKVQA (Schwenk et al., 2022). To avoid overlapping with our other KB-VQA datasets, we only use the entity set of OVEN. The entity set contains about 10K unique entities.

The original entity set contains about 5 million question-image pairs. However, the questions are highly duplicated in the original OVEN dataset. We downsample the dataset by removing repeated questions corresponding to the same entity. This reduces duplications while maximizing the diversity of the questions and coverage of entities. After the filtering, we keep 339K training samples. For validation and testing, we select 20,000 and 5,120 examples from the original OVEN Entity validation set. The original test set is not used in M2KR due to the lack of annotation.

The original task is to link the image to a specific Wikipedia Entity given a question. To formulate the task as a retrieval problem, for each entity, we use its associated Wikipedia passage as the document to retrieve. The query side of this retrieval task contains the image and its question with the inclusion of a randomly sampled instruction. Given this query, the task is to obtain the relevant Wikipedia passage. The training corpus contains about 10K passages, while the test corpus contains about 3.2K passages that cover all entities in OVEN’s original training set and validation set respectively.

A.3.3 E-VQA, Infoseek and OKVQA

E-VQA, Infoseek, and OKVQA are Knowledge-based VQA (KB-VQA) datasets. For each given image and question (with instruction), the task is to retrieve the corresponding knowledge passage. For E-VQA (Mensink et al., 2023a), the original training set contains around 1 million samples. However, it includes duplicated questions and answers referring to the same Wikipedia Entity with different query images. We filter duplicated questions that pertain to the same Wikipedia Entity. To align with the original evaluation setting of E-VQA, we further excluded samples that necessitate multiple knowledge bases, reducing the count to 167K training samples. To be consistent with the original E-VQA paper, our validation and testing sets exclusively include questions that can be answered using single knowledge. These sets contain 9,852 and 3,750 samples, respectively. We use the WikiWeb2M (Burns et al., 2023) as the knowledge source. For the training and test passage corpus, we keep all the passages that appear in the original E-VQA to align with the official E-VQA’s setting.
For **OKVQA**, we use the original training and test set. Following Lin et al. (2023b), we prepare a knowledge corpus with Wikipedia documents based on pseudo-relevance. The training and test passage corpus both contain all passages in the knowledge corpus.

For **Infoseek**, following the preprocessing steps described by Chen et al. 2023c, we use Wikipedia documents as knowledge sources and remove examples whose answers can not be found in the ground-truth documents. We randomly selected 100K examples from the training set for training and 4,708 examples from the validation set for testing (the annotation of the original test set has yet been released). The downsampling is motivated by our observation that many questions are repeated and the number of unique documents associated with the whole dataset is only about 40K. We downsampled the dataset such that the model won’t overfit severely to Infoseek passages.

Note that the aforementioned downsampling procedure for the test set is only used for constructing the M2KR benchmark. For downstream VQA evaluation, we use the same test set that existed in previous literature to ensure a fair comparison.

**Instruction templates for OVEN, Infoseek, E-VQA, and OKVQA**

- <Image> Using the provided image, obtain documents that address the subsequent question: <Question>
- <Image> Retrieve documents that provide an answer to the question alongside the image: <Question>
- <Image> Extract documents linked to the question provided in conjunction with the image: <Question>
- <Image> Utilizing the given image, obtain documents that respond to the following question: <Question>
- <Image> Using the provided image, access documents that provide insights into the following question: <Question>
- <Image> Obtain documents that correspond to the inquiry alongside the provided image: <Question>
- <Image> With the provided image, gather documents that offer a solution to the question: <Question>
- <Image> Utilizing the given image, obtain documents that respond to the following question: <Question>

## B Implementation Details

### B.1 Breakdown of Data Used in Training

In Stage 3 Full-scale Fine-tuning, the different sub-tasks in the M2KR dataset are downsampled or duplicated to balance the dataset proportions during training. The detailed breakdown of the data used in different phases is presented in Table 9. We observed that without adjusting the data proportions during training, the model’s training losses on certain datasets like WIT, Infoseek, and OKVQA decrease much faster than on others once all parameters become trainable. This goes against our goal of training a multi-tasking system. Adjusting the data proportions is crucial to ensure a more consistent learning process across different tasks.

### B.2 Detailed Hyperparameters

We use the Adam optimizer (Kingma and Ba, 2015) with a fixed learning rate of $10^{-4}$ for the mapping structure and $10^{-5}$ for the rest parameters in all experiments in all training stages. 4 Nvidia A100 GPUs were used with data parallel in all experiments.

- **Stage 0**: Training was run up to 300k steps. The batch size is 8 and the gradient accumulation step is 8. The number of negative examples is 1. The validation ran per 10k steps. The checkpoint was taken at the best Recall@50 on the original MS-MARCO validation set, following Khattab and Zaharia (2020). The total training time is approximately 1.5 days per model.

- **Stage 1**: Training was run up to 220k. The batch size is 8 and the gradient accumulation step is 8. The number of negative examples is 4. The validation interval is 10k steps. The checkpoint was taken at the best Recall@10 on the validation set of WIT in M2KR. The total training time is approximately 5 days per model.

- **Stage 2**: The intermediate pre-training was run for 12k steps for all experiments. The batch size is 8 and the gradient accumulation step is 8. The number of negative examples is 4. The total training time is approximately 5 days per model.

- **Stage 3**: Training was run for 50k for all experiments. Training was early-stopped if the performance on WIT or E-VQA decreases for 3 consecutive validation runs. Validation was run per 10k steps. The batch size is 8 and the gradient accumulation step is 8. The number of negative examples is 4. The total training time is approximately 2 days per model.
<table>
<thead>
<tr>
<th>WIT</th>
<th>IGLUE</th>
<th>KVQA</th>
<th>CC3M</th>
<th>MSMARCO</th>
</tr>
</thead>
<tbody>
<tr>
<td><img src="image1.png" alt="Image" /></td>
<td><img src="image2.png" alt="Image" /></td>
<td><img src="image3.png" alt="Image" /></td>
<td><img src="image4.png" alt="Image" /></td>
<td><img src="image5.png" alt="Image" /></td>
</tr>
<tr>
<td>Describe the image concisely.</td>
<td>Summarize the visual content of the image.</td>
<td>Provide a brief description of the image and the relevant details of the person in the image.</td>
<td>Describe the image concisely.</td>
<td>Retrieve the document that answers this question: how many years did William Bradford serve as governor of Plymouth colony?</td>
</tr>
<tr>
<td>title: PS Herald section title: Formation and operation of the North Shore Steam Company...</td>
<td>title: National Library of Uzbekistan hierarchical section title: National Library of Uzbekistan caption...</td>
<td>This is an image of Pilkington playing for Cardiff City in 2016. Anthony Pilkington date of birth is...</td>
<td>olive oil is a healthy ingredient used liberally.</td>
<td>William Bradford (c.1590 - 1657) was an English Separatist leader in Leiden,...</td>
</tr>
<tr>
<td>LLaVA</td>
<td>OVEN</td>
<td>E-VQA</td>
<td>Infoseek</td>
<td>OKVQA</td>
</tr>
<tr>
<td><img src="image6.png" alt="Image" /></td>
<td><img src="image7.png" alt="Image" /></td>
<td><img src="image8.png" alt="Image" /></td>
<td><img src="image9.png" alt="Image" /></td>
<td><img src="image10.png" alt="Image" /></td>
</tr>
<tr>
<td>Provide a brief description of the image along with the following question: what unique situation is occurring in this soccer match?</td>
<td>Using the provided image, obtain documents that address the subsequent question: what is this park called?</td>
<td>Obtain documents that correspond to the inquiry alongside the provided image: how big can this plant become?</td>
<td>With the provided image, gather documents that offer a solution to the question: What is the country of origin of this food?</td>
<td>Using the provided image, obtain documents that address the subsequent question: How many teeth does this animal use to have?</td>
</tr>
<tr>
<td>In this soccer match, a unique situation is occurring where three men are playing against each other, each wearing a different colored uniform.</td>
<td>Nationals Park is a baseball stadium along the Anacostia River in the Navy Yard neighborhood...</td>
<td>Dwarf cornel is a rhizomatous herbaceous perennial growing to 20cm (8 inches) tall...</td>
<td>title: Submarine sandwich content: Submarine sandwich A submarine sandwich, also known as a sub...</td>
<td>Most cats have 26 deciduous teeth and 30 permanent teeth.</td>
</tr>
</tbody>
</table>

Table 8: Demonstration of the retrieval tasks for each dataset. We show the image (first row) query, the text query (second row), and the retrieved ground truth document (third row) for each dataset. Since some retrieved documents are too long, we only show part of the document and use ... to stand for continuing documents.
<table>
<thead>
<tr>
<th></th>
<th>Stage 1</th>
<th>Stage 2</th>
<th>Stage 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>WIT</td>
<td>2.8M</td>
<td></td>
<td>140K</td>
</tr>
<tr>
<td>IGLUE</td>
<td></td>
<td></td>
<td>-</td>
</tr>
<tr>
<td>KVQA</td>
<td>65K</td>
<td></td>
<td>6.5K</td>
</tr>
<tr>
<td>CC3M</td>
<td>595K</td>
<td></td>
<td>29.8K</td>
</tr>
<tr>
<td>MSMARCO</td>
<td>400K</td>
<td></td>
<td>40K</td>
</tr>
<tr>
<td>OVEN</td>
<td>339K</td>
<td></td>
<td>33.9K</td>
</tr>
<tr>
<td>LLaVA</td>
<td>351K</td>
<td></td>
<td>35.1K</td>
</tr>
<tr>
<td>OKVQA</td>
<td>9K</td>
<td></td>
<td>90K (repeat 10 times)</td>
</tr>
<tr>
<td>Infoseek</td>
<td>100K</td>
<td></td>
<td>50K</td>
</tr>
<tr>
<td>E-VQA</td>
<td>167K</td>
<td>167K</td>
<td>167K</td>
</tr>
</tbody>
</table>

Table 9: The dataset sizes are adjusted in Stage 3 in practice.

Single-task Downstream Fine-tuning: The batch size is 8 and the gradient accumulation step is 8. The number of negative examples is 4. For reference, in our experiments, the downstream fine-tuning took 20k, 5k, 1k, 15k, 2.5k steps to achieve the best performance for WIT, OVEN, Infoseek, E-VQA, OKVQA respectively (for the ViT-G + Base-v2 PreFLMR). The total training time is approximately 1 days per model per task.

VQA Fine-tuning: We used BLIP2-T5XL as the answer generator as in RA-VQAav2 (Lin et al., 2023b). The retriever was frozen during training and inference. The batch size is 1 and the gradient accumulation step is 16. For each question in a training batch, top-5 relevant documents were pre-extracted using the retriever, and 3 out of 5 were randomly selected. These 3 documents were concatenated to the question and sent to the answer generator for one forward pass individually. This setting is to enable training with top-5 documents given limited GPU memory. The total training time is approximately 2 days per model.

Every model reported in this paper was reproduced once to make sure the training is reproducible. The best result is reported since the model with the best result will be released to the community. There is not much difference in the two runs. The absolute difference is less than 0.2 Recall score in most datasets (except that PreFLMR_ViT-B_Base-v2 has a -0.4 difference on Infoseek).

B.3 Large-v1 Training

In our experiments, we found that training Large-v1 during Stage 2/3 was not steady. First, the loss decreased faster than in other systems, like Base-v1, even though Large-v1 had worse system performance. This happened because Large-v1’s bigger model capacity made it more prone to overfitting.

Next, the loss suddenly shot up, causing the training to collapse, despite using the same data and strategy as Base-v1. We tried different hyperparameters, like lowering the learning rate to $1e^{-6}$, $3e^{-6}$, but the model still collapsed.

Finally, when we used LoRA (Hu et al., 2022) with Large-v1 during training, it helped stabilize the process. The LoRA hyperparameters used were: $\gamma = 16$, $\alpha = 32$, and a dropout rate of 0.05.

B.4 Model Design in Detail

Similar to FLMR, PreFLMR consists of three components: a vision model $\mathcal{F}_V$, a mapping structure $\mathcal{F}_M$, and a language model $\mathcal{F}_L$.

Feature Extraction. The textual query $q$ consists of an instruction and (optionally) a question (e.g., "Utilize the given image to procure documents addressing the following query: [Question]"). We use a language model with hidden size $d_L$ to obtain embeddings for all $N_q$ tokens which are concatenated into matrix $Q_q$:

$$Q_q = \mathcal{F}_L(q) \in \mathbb{R}^{N_q \times d_L}$$

Like FLMR, a vision model $\mathcal{F}_V$ encodes the input image $I$, extracting the [CLS] token embeddings from the last layer. PreFLMR additionally uses the patch embeddings from the penultimate layer of ViT for more complete representation.

$$Q_{I,[CLS]} = \mathcal{F}_V(I) \in \mathbb{R}^{1 \times d_V}$$

The mapping structure $\mathcal{F}_M$ comprises two components: a 2-layer MLP $\mathcal{F}_M^{MLP}$ and a Transformer block $\mathcal{F}_M^{TR}$.

Following the FLMR model, a 2-layer Multi-Layer Perceptron (MLP) $\mathcal{F}_M^{MLP}$ is utilized to convert the initial token embeddings into visual token embeddings with a length of $N_v$ and a hidden size $d_h$:

$$Q_{I}^{MLP} = \mathcal{F}_M^{MLP}(Q_{I,[CLS]}) \in \mathbb{R}^{N_v \times d_h}$$

Moreover, an additional Transformer module $\mathcal{F}_M^{TR}$ is introduced to manage all patch embeddings. It is a stack of $N_{TR}$ transformer layers with a hidden size $d_L$, followed by a simple MLP layer at the

$$\mathcal{F}_M^{TR} : \mathbb{R}^{d_V} \rightarrow \mathbb{R}^{N_v \times d_h / 2} \rightarrow \mathbb{R}^{N_v \times d_h}$$, subsequently reshaped into $\mathbb{R}^{N_v \times d_h}$.
end. This module leverages cross-attention with the text query $Q_q$, enabling query-aware image feature mapping.

$$Q^{TR}_I = F^R_M(F_V(Q_{I,PATCH}, Q_q)) \in \mathcal{R}^{N_V \times d_h}$$

(6)

Here, $F_V$ represents a 1-layer MLP that adapts the dimension from $d_V$ to $d_L$, which is subsequently transformed to $d_h$ by the linear MLP layer of $F^R_M$. The resultant features from these processes are concatenated to formulate the query embeddings:

$$Q = [Q_q]Q^{MLP}_I Q^{TR}_I \in \mathcal{R}^{(N_v + N_V + N_q) \times d_h}$$

(7)

Furthermore, the document representations in the knowledge base are denoted by $D$, derived from the document content $d$ with length $l_D$:

$$D = F_I(F_L(d)) \in \mathcal{R}^{l_D \times d_h},$$

(8)

where $F_I$ signifies a straightforward MLP layer tasked with mapping $d_L$ to $d_h$, thereby aligning the dimensionality with the query embeddings.

**Multi-Modal Late Interaction.** The relevance score between a question-image pair $\bar{q} = (q, I)$ and a document $d$ is calculated using a late-interaction paradigm:

$$r(\bar{q}, d) = r((q, I), d) = \sum_{i=1}^{l_Q} \max_{j=1}^{l_D} Q_{i} D_{j}^\top$$

(9)

where $l_Q = N_v + N_V + N_q$. For each token in the query, the system aggregates the maximum relevance score across all tokens in the document.

**Training and Inference.** For model training, documents $d^*$ corresponding to a query $q$ are considered gold (positive) samples. We incorporate random negative sampling from the corpus. Additionally, we adopt in-batch negative sampling as suggested by Karpuikhin et al. (2020), treating all non-corresponding documents in a batch as negatives for $q$, denoted as $N(q)$. The model is trained using a contrastive loss across the dataset $D$:

$$\mathcal{L} = - \sum_{(q, d^*) \in D} \log \frac{\exp(r(q, d^*))}{\exp(r(q, d^*)) + \sum_{z \in N(q)} \exp(r(q, z))}$$

(10)

Post-training, all documents are indexed through PLAID (Santhanam et al., 2022a) for efficient late-interaction retrieval. For detailed evaluation of retrieval efficiency, we refer readers to Lin et al. (2023b).

**C Ablation Study on Pre-training Stages**

We present the ablation study for the four pre-training stages in Table 10. To ensure consistent comparison, these ablated versions underwent the same number of training steps as PreFLMR_ViT-B_Base-v2. The results clearly indicate that the removal of any stage deteriorates performance. Specifically, disabling Stage 0 (i.e., using untrained text encoder) leads to the most significant performance decline because the text encoder is not pre-trained on late-interaction, resulting in a diminished ability to capture fine-grained relevance within the same computational budget. Note that removing Stage 0 leads to collapsed performance on Stage 1, where the text encoder is frozen. Furthermore, removing Stage 2 notably affects the performance on E-VQA more than on other KB-VQA datasets, highlighting the challenge posed by E-VQA and the necessity of intermediate pre-training.

**D V-Entropy-based Analysis of Intermediate Pre-training**

V-Entropy (Xu et al., 2020), $H_Y(Y | X)$, is the minimal Negative Log-Likelihood (NLL) achievable by the probabilistic predictor $f(Y | X)$ under the predicative family $\mathcal{V}$. A predicative family can be viewed as the set of reachable models under a certain model architecture and training budgets.

We define Mutual Information $I_{Y|N_f}(D_1 \rightarrow D_2)$ between datasets $D_1$ and $D_2$ in Eq.11. We define $H_{Y|N_f}(D_2)$ as the minimal achieved NLL loss on the validation set of dataset $D_2$ after $N_f$ training steps on $D_2$. $\mathcal{V}[N_f, D_1, N_1]$ denotes the set of reachable models after $N_f$ fine-tuning steps on $D_2$ starting from a checkpoint that has been trained on dataset $D_1$ for $N_1$ steps. This is V-Entropy with additional predicative family specification.

$$I_{Y|N_f}(D_1 \rightarrow D_2) = H_{Y|N_f}(D_2) - H_{Y|N_f,D_1,N_1}(D_2)$$

(11)

Intuitively, $D_1$ has high mutual information with $D_2$ if models initialized from $D_1$ checkpoints attain much lower NLL loss compared to models...
initialized without training on $D_1$. $N_f$ and $N_t$ set the computation constraints for training on $D_2$ and $D_1$, respectively. In our experiment, $V$ is the Pre-FLMR architecture, $D_1$ is the E-VQA dataset and $D_2$ is the training set of M2KR. $N_f$ corresponds to $N_{inter}$ in Sec. 5.6, which is the intermediate training steps on the E-VQA dataset. In the analysis, we set $N_f$ to 5,000 and sweep $N_t$ from 0 to 25,000 in intervals of 5,000.

We refer readers to Xu et al. (2020) for detailed properties of V-Entropy and emphasize that $I_{V[N_f]}(D_1 \rightarrow D_2)$ is an empirical value we define to estimate mutual information between datasets. It is different from the V-Information defined in Xu et al. (2020) which estimates the mutual information between model input and output.

### E Qualitative Analysis for OKVQA and E-VQA

In this section, we compare examples from the OKVQA and E-VQA datasets to highlight their differences. To avoid cherry-picking, we use examples from its official website\(^9\) for OKVQA. Similarly, we use the examples included in the paper for E-VQA. Table 11 presents three examples from each dataset.

The OKVQA examples typically require common sense knowledge, like ‘people attend church on Sundays’ or ‘firetrucks use fire hydrants.’ State-of-the-art Large Language Models (LLMs) often have this common sense knowledge inherently built-in, making additional knowledge retrieval less impactful for OKVQA tasks.

In contrast, E-VQA examples demand more specialized, expert-level knowledge, necessitating an effective knowledge retrieval system. For instance, correctly answering a question about 'Acacia paradoxa' requires first retrieving the relevant document providing specific information about this plant species. Enhancing the knowledge retrieval system to source accurate documents is crucial for improving performance on the E-VQA dataset.

### F Artifacts and License

We list the resources used and their License below:

1. huggingface-transformers (Apache License 2.0) provides pre-trained model checkpoints for BLIP 2, DPR, and their tokenizers: https://github.com/huggingface/transformers
2. FAISS (Johnson et al., 2019) (MIT License) is used to index document embeddings for fast retrieval with DPR: https://github.com/facebookresearch/faiss
3. huggingface-PEFT (Apache License 2.0) for parameter-efficient LoRA fine-tuning: https://github.com/huggingface/peft
4. PLAID and ColBERTv2 (MIT License): https://github.com/stanford-futuredata/ColBERT
6. Datasets used in building the M2KR benchmark:
   - WIT (Creative Commons Attribution-ShareAlike 3.0 Unported https://github.com/google-research-datasets/wit/blob/main/LICENSE);
   - MSMARCO (non-commercial research purposes only https://microsoft.github.io/msmarco/);
   - CC3M (Free for any purposes only https://microsoft.github.io/cc3m);
   - LLaVA, the image of LLaVA is a subset of CC3M. It should inherit the license of CC3M. The conversation data follows policy of OpenAI: https://openai.com/policies/terms-of-use.
   - IGLUE (MIT license https://github.com/eebug/iglue/blob/main/LICENSE);

---
\(^9\)https://okvqa.allenai.org/

<table>
<thead>
<tr>
<th>Model</th>
<th>WIT</th>
<th>IGLUE</th>
<th>KVQA</th>
<th>MM</th>
<th>OVEN</th>
<th>LLaVA</th>
<th>Infoseek</th>
<th>E-VQA</th>
<th>OKVQA</th>
</tr>
</thead>
<tbody>
<tr>
<td>PreFLMR_ViT-B_Base-v1</td>
<td>41.7</td>
<td>57.3</td>
<td>28.6</td>
<td>79.5</td>
<td>46.3</td>
<td>67.2</td>
<td>48.8</td>
<td>67.9</td>
<td>66.1</td>
</tr>
<tr>
<td>w/o Stage 0</td>
<td>25.5</td>
<td>28.8</td>
<td>21.0</td>
<td>56.5</td>
<td>33.9</td>
<td>55.0</td>
<td>42.5</td>
<td>51.8</td>
<td>64.5</td>
</tr>
<tr>
<td>w/o Stage 1</td>
<td>38.2</td>
<td>54.9</td>
<td>26.6</td>
<td>78.0</td>
<td>45.5</td>
<td>62.8</td>
<td>44.6</td>
<td>61.9</td>
<td>65.5</td>
</tr>
<tr>
<td>w/o Stage 2</td>
<td>41.2</td>
<td>56.8</td>
<td>26.5</td>
<td>78.2</td>
<td>43.7</td>
<td>65.0</td>
<td>47.0</td>
<td>57.3</td>
<td>65.1</td>
</tr>
</tbody>
</table>
Table 11: Demonstrative examples from OKVQA and E-VQA. Questions in E-VQA require more domain knowledge to answer generally.

- KVQA (No specific license is mentioned https://mallabiisc.github.io/resources/kvqa/);
- OVEN (Apache-2.0 license https://github.com/open-vision-language/oven/blob/main/LICENSE);
- E-VQA (no specific license mentioned https://github.com/google-research/google-research/tree/master/encyclopedic_vqa);
- Infoseek (Apache License 2.0 https://github.com/open-vision-language/infoseek/blob/main/LICENSE);
- OKVQA (Copyright (c) 2021, Chen Qu and Center for Intelligent Information Retrieval, University of Massachusetts, Amherst. https://github.com/prdwb/okvqa-release/blob/main/LICENSE)

G PreFLMR model performance radar chart on M2KR tasks

Fig. 3 demonstrates the performance of PreFLMR with a radar plot. The best and worst numbers of each task are annotated.

H AI Assistance

Our coding work was assisted by Github Copilot.10 OpenAI ChatGPT11 was only used in proofreading and spell-checking. We claim that the content presented in this paper was fully original.

In particular, we emphasize that no changes are made to the original data of all the datasets used in our work. Our released models and artifacts should only be used for non-commercial purposes. By using the pre-trained models, users agree to respect the terms and conditions of the datasets used in pre-training.

10https://github.com/features/copilot
11https://chat.openai.com/
Figure 3: PreFLMR achieves strong performance on the M2KR benchmark. The scale of the plot is adjusted for better visualization. The best and worst numbers of each task are annotated.