MM-EMBED: UNIVERSAL MULTIMODAL RETRIEVAL WITH MULTIMODAL LLMS

Sheng-Chieh Lin *[1],[2] **Chankyu Lee** [1] **Mohammad Shoeybi** [1] **Jimmy Lin** [2]

Bryan Catanzaro [1] **Wei Ping** * [1]

[1] NVIDIA [2] University of Waterloo

ABSTRACT

State-of-the-art retrieval models typically address a straightforward search scenario, where retrieval tasks are fixed (e.g., finding a passage to answer a specific question) and only a single modality is supported for both queries and retrieved results. This paper introduces techniques for advancing information retrieval with multimodal large language models (MLLMs), enabling a broader search scenario, termed universal multimodal retrieval, where multiple modalities and diverse retrieval tasks are accommodated. To this end, we first study fine-tuning an MLLM as a bi-encoder retriever on 10 datasets with 16 retrieval tasks. Our empirical results show that the fine-tuned MLLM retriever is capable of understanding challenging queries, composed of both text and image, but underperforms a smaller CLIP retriever in cross-modal retrieval tasks due to *modality bias* from MLLMs. To address the issue, we propose modality-aware hard negative mining to mitigate the *modality bias* exhibited by MLLM retrievers. Second, we propose to continually fine-tune the universal multimodal retriever to enhance its text retrieval capability while maintaining multimodal retrieval capability. As a result, our model, MM-Embed, achieves state-of-the-art performance on the multimodal retrieval benchmark M-BEIR, which spans multiple domains and tasks, while also surpassing the state-of-the-art text retrieval model, NV-Embed-v1, on MTEB retrieval benchmark. Finally, we explore to prompt the off-the-shelf MLLMs as the zero-shot rerankers to refine the ranking of the candidates from the multimodal retriever. We find that through prompt-and-reranking, MLLMs can further improve multimodal retrieval when the user queries (e.g., text-image composed queries) are more complex and challenging to understand. These findings also pave the way to advance universal multimodal retrieval in the future. We release the model weights at: `https://huggingface.co/nvidia/MM-Embed`.

1 INTRODUCTION

Information retrieval is crucial for a variety of downstream tasks, such as question answering (Kwiatkowski et al., 2019), fact-checking (Thorne et al., 2018), and retrieval-augmented generation (Lewis et al., 2020). Existing state-of-the-art retrievers often focus on narrow scenarios. For example, LLM-based retrievers (Wang et al., 2023; Lee et al., 2024; Meng et al., 2024; Moreira et al., 2024) are limited to text-to-text retrieval tasks, where both the query and the retrieved results are text-only. Recent work on multimodal retrieval (Zhang et al., 2024; Jiang et al., 2024) focuses on specific tasks and assumes a homogeneous document format. However, in real-world applications, documents and queries often consist of diverse formats or modalities, such as text, images, and interleaved text and images. To advance information retrieval and support broader search scenarios, this work explores the use of multimodal LLMs (MLLMs; Dai et al., 2024; Liu et al., 2023a; 2024) for universal multimodal retrieval, accommodating diverse user-instructed tasks with multimodal queries and documents, as illustrated in Figure 1.

*Sheng-Chieh Lin did this work during an internship at NVIDIA. Correspondence to: Sheng-Chieh Lin ⟨s269lin@uwaterloo.ca⟩, Wei Ping ⟨wping@nvidia.com⟩.

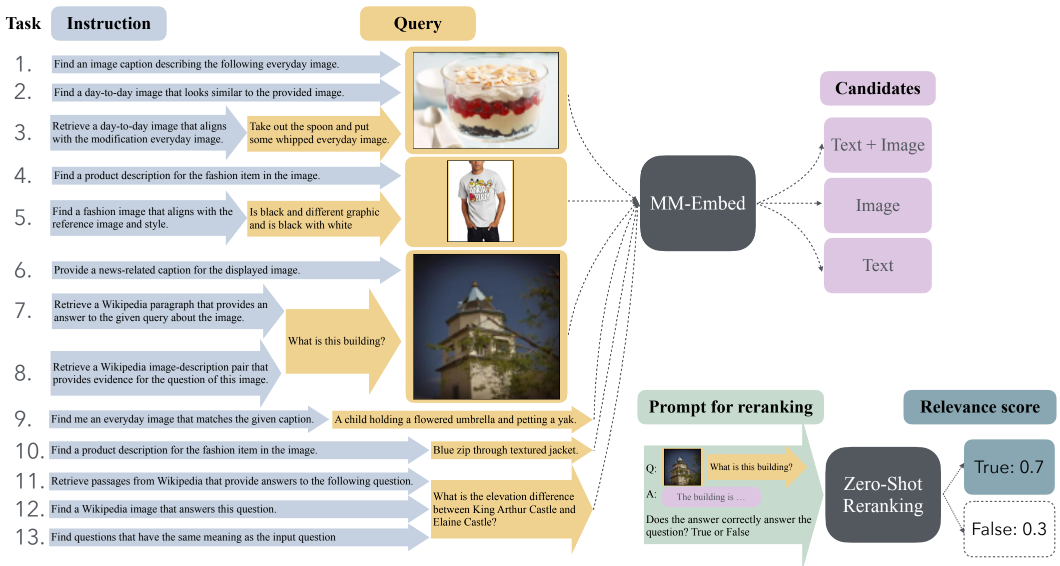


Figure 1: Illustration of universal multimodal retrieval, where diverse tasks with instructions, queries and documents with multimodal formats are supported. In this work, we explore to fine-tune MLLM-based universal multimodal retriever, MM-Embed, and prompt an MLLM for reranking.

We first explore to fine-tune MLLM-based bi-encoder retrievers with instructions as a guide (Asai et al., 2023) on 16 multimodal retrieval tasks from M-BIER (Wei et al., 2023). We find that MLLM-based retrievers significantly outperform CLIP-based retrievers in the challenging tasks, where interleaved text–image queries are given, such as visual question answering and composed image retrieval (tasks 3 and 7 in Figure 1). However, MLLM-based retrievers underperform in cross-modal retrieval tasks due to the *modality bias* from MLLMs. That is, given a text-based query with the instruction to retrieve an image (e.g., task 9 in Figure 1), an MLLM-based retriever tends to retrieve a relevant text-only rather than documents with images, especially when we improve the MLLM-based retriever's text retrieval capability. To address the issue, we propose modality-aware hard negative mining in Section 4.1.1 and continual text-to-text retrieval fine-tuning in Section 4.1.2. Our final retriever, coined MM-Embed, is the first state-of-the-art universal multimodal retriever while maintaining competitive text-to-text retrieval performance across diverse tasks.

Finally, we explore to prompt MLLMs as zero-shot rerankers. Surprisingly, we find that the zero-shot MLLM-based rerankers can further boost retrieval accuracy in the tasks, where user queries are interleaved text–image and more challenging to understand. For example, in the composed image retrieval dataset, CIRCO (Baldrati et al., 2023), the zero-shot rerankers are able to refine the ranked lists and significantly boosts the accuracy (mAP@5) over 7 points from the existing state-of-the-art composed-image retriever (Zhang et al., 2024) and our universal multimodal retrievers. This finding indicates that there is still room for improvement in such challenging tasks in order to tackle universal multimodal retrieval. Also, knowledge distillation from zero-shot or few-shot MLLM-based rerankers to retrievers is a promising direction.

We summarize our contributions as follows: *i)* We present a study on applying MLLMs to universal multimodal retrieval. *ii)* We are the first to build MLLM-based universal multimodal retrievers. Notably, our MM-Embed, initialized from the existing best-performing text retriever (NV-Embed-v1; Lee et al., 2024), not only achieves state-of-the-art results in universal multimodal retrieval benchmark, M-BEIR (Wei et al., 2023), but also surpasses NV-Embed-v1 in text-to-text retrieval tasks on MTEB. *iii)* We are the first work to explore prompting MLLMs for zero-shot reranking. With a zero-shot MLLM-based reranker, we are able to boost the ranking accuracy over 7 points upon state-of-the-art retrievers in the composed image retrieval task, CIRCO (Baldrati et al., 2023).

We organize the rest of the paper as follows. We discuss related work in § 2. We introduce the definition of universal multimodal retrieval in § 3 and present the proposed method in § 4. We report experiment results in § 5 and conclude the paper in § 6.

2 RELATED WORK

Instruction-Aware Dense Representation Learning. Asai et al. (2023) is the first work to identify the implicit search intent behind each retrieval task and propose to fine-tune a retriever to learn diverse retrieval tasks with handwritten task instructions. Su et al. (2023) and existing state-of-the-art LLM-based text embedding models (Wang et al., 2023; Meng et al., 2024; Lee et al., 2024) adopt this approach to broader tasks beyond text retrieval, such as text classification and clustering. Recently, Wei et al. (2023) propose a universal multimodal retrieval dataset, M-BEIR, and find that instruction-aware dense retrieval fine-tuning is crucial to tackle universal multimodal retrieval.

Vision-Language Models for Multimodal Retrieval. With the advance of pre-trained vision-language models (Radford et al., 2021; Li et al., 2022), research focus shifts from single-modal (Bajaj et al., 2016; Fu et al., 2023) to cross-modal (Lin et al., 2014; Han et al., 2017; Liu et al., 2021a) or more complex multimodal retrieval tasks (Liu et al., 2021b; Wu et al., 2021; Baldrati et al., 2023). However, the aforementioned tasks assume homogeneous modality for queries and documents, limiting its application. Liu et al. (2023c) take one step further to tackle the retrieval scenario involving candidate pool with heterogeneous modalities but still limit to single retrieval task.

Wei et al. (2023) extend the study to a more general scenario, where retrievers are required to deal with queries, candidate pool in heterogeneous modalities and diverse retrieval tasks. However, the study is limited to CLIP-based retrievers and ignores important text-to-text retrieval tasks, such as fact checking (Thorne et al., 2018) and entity retrieval (Hasibi et al., 2017). While Koukounas et al. (2024) aim to fine-tune a CLIP-based retriever with both strong text-to-text and multimodal retrieval capability, they only consider simple multimodal retrieval tasks: image-caption retrieval (Young et al., 2014; Lin et al., 2014). Concurrent to our work, Jiang et al. (2024) propose to fine-tune MLLMs on NLI dataset (Bowman et al., 2015) and demonstrate their transferability to multimodal retrieval. In this paper, we are the first to study how to fine-tune an MLLM-based universal multimodal retriever while maintaining strong text-to-text retrieval capability. Also, we are the first to explore prompting MLLMs as zero-shot rerankers in diverse multimodal retrieval tasks.

3 UNIVERSAL MULTIMODAL RETRIEVAL

Following the framework of Lin et al. (2021), we formulate the task of retrieval as follows: given a query q, the goal is to retrieve a ranked list of candidates $\{c_1, c_2, \cdots c_k\} \in C$ to maximize some ranking metrics, such as nDCG, where C is the collection of documents. In this work, we borrow the setting of universal multimodal retrieval from Wei et al. (2023), where user queries and candidates may consist of a text, image or interleaved text–image; i.e., $q \in \{q^{\text{txt}}, q^{\text{img}}, (q^{\text{txt}}, q^{\text{img}})\}$; $c \in \{c^{\text{txt}}, c^{\text{img}}, (c^{\text{txt}}, c^{\text{img}})\}$. Additionally, there are multiple search intents behind a search query, which can be elaborated by task-specific instructions (Asai et al., 2023). For example, in task 1 and 2 of Figure 1, given the same image as a query, the search intent is to find an image caption and similar image, respectively. Thus, in universal multimodal retrieval, given a multimodal query and task instruction $inst$, we aim to retrieve a list of candidates from a pool of multimodal documents to maximize a specified ranking metric. Note that we only consider text and image in this work while more modalities, such as audio and video can be included, which we leave for future work.

4 METHOD

In this section, we describe our approach to universal multimodal retrieval by leveraging multimodal LLMs (MLLMs). In Section 4.1, we first fine-tune an MLLM-based retriever to project multimodal user queries, along with task instructions, into the same semantic space as multimodal documents, enabling k-nearest neighbor search (Johnson et al., 2021). In Section 4.2, we present our method for using MLLMs to rerank the top-k candidates retrieved by the universal multimodal retriever.

4.1 FINE-TUNING MULTIMODAL LLMS FOR UNIVERSAL MULTIMODAL RETRIEVAL

We fine-tune an MLLM-based retriever parameterized by θ (i.e., η^{θ}) under the guidance of task-specific instructions, aiming to capture the implicit intents behind retrieval tasks. Specifically, given a user query q_i with the specified task instruction $inst_i$ and its relevant and negative candidates, c_i^+ and c_i^-, we minimize the InfoNCE loss (Gutmann & Hyvärinen, 2010):

$$NCE = -\frac{1}{|\mathcal{B}|}\sum_{i=1}^{|\mathcal{B}|}\log\frac{\exp\left(\eta^\theta(inst_i, q_i)\cdot\eta^\theta(c_i^+)/\tau\right)}{\sum_{c'\in\mathcal{D}_\mathcal{B}}\exp(\eta^\theta(inst_i, q_i)\cdot\eta^\theta(c')/\tau)}, \tag{1}$$

where $\mathcal{D}_\mathcal{B} = (c_1^+, c_1^-, \cdots, c_{|\mathcal{B}|}^+, c_{|\mathcal{B}|}^-)$ includes all the positive and negative documents for all the queries in the mini batch $\mathcal{B}$, $\eta^\theta(\cdot) \in \mathbb{R}^d$ is a normalized vector and τ is the temperature.

4.1.1 Modality-Aware Hard Negative Mining

Prior work (Karpukhin et al., 2020; Xiong et al., 2021; de Souza P. Moreira et al., 2024) has demonstrated that hard negative mining significantly improves representation learning for text-to-text retrieval. In the previous retrieval setting, where the corpus consists of documents with a homogeneous modality, a document is considered a hard negative if it lacks the required information but is still retrieved by a model. However, in the scenario of universal multimodal retrieval, where the corpus contains documents involving diverse modalities, the users' desired modality as specified in task instructions (i.e., text, image or interleaved text–image) should be taken into consideration. For example, as shown in Figure 1, the first and second users issue the same query along with different instructions, requiring the documents to be in the format of text and image, respectively. To address this, we propose modality-aware hard negative mining to guide models in retrieving candidates that meet both the users' information needs and their preferred modality.

Specifically, we first fine-tune an MLLM-based retriever using random negatives; i.e., $\mathcal{D}_\mathcal{B} = (c_1^+, \cdots, c_{|\mathcal{B}|}^+)$. The fine-tuned model is denoted as M^{rand}. For each query q_i and its associated instruction $inst_i$ in the training set, we generate two types of negatives from the top-50 candidates retrieved by M^{rand}: *i*) negatives with incorrect modality (C_i^1), where the candidate ranks higher than the labeled positive but has a different modality from the desired one, and *ii*) negatives with unsatisfactory information (C_i^2), where the candidate ranks lower than k' but has the same desired modality. Note that setting k' to a small number may include false positives while setting k' to large number would make the negative samples too easy. Thus, in our experiment, following the prior work (Chen et al., 2022; Lin et al., 2023), we set $k' = 45$. While training, given the query q_i with the associated instruction $inst_i$, we generate a triplet, $((inst_i, q_i), c_i^+, c_i^-)$, by sampling hard negative c_i^- from either C_i^1 or C_i^2 with the same probability. We denote the models fine-tuned with modality-aware hard negatives as M^{hard}. We refer readers to Fig. 2 in the Appendix for examples of both types of negative samples.

4.1.2 Continual Text-to-Text Retrieval Fine-Tuning

Since text-to-text retrieval remains one of the most commonly used retrieval tasks, we further fine-tune M^{hard} on diverse public text-to-text retrieval tasks, including MS MARCO (Bajaj et al., 2016), HotpotQA (Yang et al., 2018), Natural Question (Kwiatkowski et al., 2019), PAQ (Lewis et al., 2021), StackExchange (Stack-Exchange-Community, 2023), Natural Language Inference (Bowman et al., 2015), SQuAD (Rajpurkar et al., 2016), ArguAna (Wachsmuth et al., 2018), BioASQ (Nentidis et al., 2023), FiQA (Maia et al., 2018), and FEVER (Thorne et al., 2018). As these datasets do not contain negative samples, we employ the fine-tuned LLM-based retriever (NV-Embed-v1; Lee et al., 2024) to mine hard negatives in our experiments (see de Souza P. Moreira et al. (2024) for details).

During the continual fine-tuning stage, we uniformly sample triplets from both the universal multimodal and text-to-text retrieval training data. Note that for each query q_i in universal multimodal retrieval training data, we use M^{hard} to mine the second-type hard negatives C_i^2 again. Since no first-type hard negatives (i.e., $C_i^1 = \emptyset$) are mined by M^{hard}, we retain the first-type hard negative mined by M^{rand}.

4.2 Prompting Multimodal LLMs for Reranking

Prior work (Sun et al., 2023; Jin et al., 2024) has demonstrated that instruction fine-tuned LLMs can be prompted to rerank candidates in text-to-text retrieval tasks. In this work, we prompt

LLaVa-Next (Liu et al., 2024) to further rerank the top-10 retrieved candidates by universal multimodal retrievers. Following the approach in Nogueira et al. (2020), we frame the reranking task as a series of true-false questions. Specifically, given a query and retrieved candidate, we prompt LLaVa-Next to determine whether the retrieved candidate satisfies the given query by answering "True" or "False". For example, in the image caption retrieval (task 1 in Figure 1), given an image query, q^{img}, and a retrieved text-based candidate, c^{txt}, we use the below prompt: "$< q^{img} >$*\nCaption:*$< c^{txt} >$*\nDoes the above daily life image match the caption? True or False*". Additionally, in the visual question answering retrieval (task 7 in Figure 1), given a visual question, *<Qry image><Qry text>*, and a retrieved text-based candidate, *<Doc text>*, we use the below prompt: *<Qry image>\nQuestion:<Qry text>\nAnswer:<Doc text>\nDoes the answer correctly answer the question? True or False*. We refer readers to Table 13 in the Appendix for the specific prompts used in different multimodal retrieval tasks.

To compute relevance scores, we apply the `Softmax` operation over the logits of the "True" and "False" tokens, using the probability of the "True" token as the relevance score for reranking. Our preliminary study in Section 5.3.3 shows that zero-shot MLLM-based rerankers mainly improve the tasks, where queries are interleaved text–image, such as composed image retrieval and visual question answering as shown in the tasks 3, 5,7 and 8 of Figure 1.

5 Experiments

5.1 Datasets and Models

Multimodal Retrieval Dataset. We evaluate models' universal multimodal retrieval capability using M-BEIR dataset (Wei et al., 2023), which is constructed from 10 datasets with 16 diverse multimodal retrieval tasks across 4 domains listed in Table 8 (in the Appendix).[1] We train our models on the M-BEIR 1.1M training queries and evaluate models' effectiveness on the 190K test queries. Following the global evaluation setting of M-BEIR dataset, for each query, candidates are retrieved from a merged candidate pool of 5.6M multimodal documents spanning all 10 datasets. We report the averaged Recall@5 (R@5) as retrieval accuracy across all test queries in each dataset, except for Fashion200K and FashionIQ, where we report Recall@10 (R@10). We refer readers to Wei et al. (2023) for more details on the construction of M-BEIR dataset.

Text-to-Text Retrieval Dataset. While M-BEIR contains WebQA dataset for text-to-text retrieval evaluation, we conduct a more comprehensive text-to-text retrieval evaluation using MTEB dataset (Muennighoff et al., 2023). Specifically, we evaluate our models on 15 diverse text retrieval datasets.[2] Following the established procedure, we report the averaged nDCG@10 across the 15 text retrieval datasets. Note that unlike in M-BEIR, where candidates are retrieved from a merged pool across all tasks, in the MTEB retrieval tasks, we retrieve candidates from separate corpora for each task.

Backbone Model Choices. In this work, we utilize two representative backbones of vision–language models to build universal multimodal retrievers, CLIP (Radford et al., 2021) and LLaVa-Next (Liu et al., 2024). For CLIP, we initialize from CLIP-large model and employ the best-performing modeling approach from Wei et al. (2023), denoted as CLIP_{SF}.[3] This method fuses input image and text features by separately encoding each input (query or document) image and text into separate vectors, which are then summed to create a fused vector (Liu et al., 2023c).

LLaVa-Next (Liu et al., 2024) is a multimodal LLM (MLLM), which integrates a CLIP image encoder, LLM and a vision–language MLP projector to align image features to the input embedding space of the LLM. We use LLaVa-Next with Mistral 7B (Jiang et al., 2023) as the backbone LLM.[4] We experiment with three variants: (1) LLaVa-E: the *<eos>* token embedding is used to aggregate information from the multimodal input, a method commonly employed in prior work for text retrieval (Wang et al., 2023; Ma et al., 2024b); (2) LLaVa-P: the MLLM is prompted to summarize

[1] https://huggingface.co/datasets/TIGER-Lab/M-BEIR

[2] The 15 retrieval datasets in MTEB are derived from public datasets in BEIR (Thakur et al., 2021), excluding BioASQ, Signal-1M, TREC-NEWS, Robust04.

[3] https://huggingface.co/openai/clip-vit-large-patch14

[4] https://huggingface.co/llava-hf/llava-v1.6-mistral-7b-hf

Table 1: Main results. Following Wei et al. (2023), we report R@5 for all the datasets, except for Fashion200K and FashionIQ, where we report R@10. The tasks of single-modal and multi-modal queries denote tasks 1–5 and 6–8, respectively. For MTEB text retrieval (Muennighoff et al., 2023), we report nDCG@10 averaged from 15 retrieval tasks (detailed in Appendix Table 9).

Task	**Dataset**	M^{rand}				M^{hard}			MM-Embed
		CLIP$_{\text{SF}}$	LLaVa-E	LLaVa-P	NV-Embed-v1	CLIP$_{\text{SF}}$	LLaVa-P	NV-Embed-v1	
1. $q^{\text{txt}} \to c^{\text{img}}$	VisualNews	43.8	33.2	34.2	32.1	42.7	39.7	41.1	41.0
	MSCOCO	72.0	69.3	70.8	64.6	69.2	73.8	72.7	71.3
	Fashion200K	16.4	13.5	13.3	10.4	19.7	17.4	18.6	17.1
2. $q^{\text{txt}} \to c^{\text{txt}}$	WebQA	83.2	88.6	88.8	92.1	88.2	93.6	95.6	95.9
3. $q^{\text{txt}} \to (c^{\text{img}}, c^{\text{txt}})$	EDIS	46.5	55.9	56.6	55.1	54.2	68.8	69.8	68.8
	WebQA	76.0	80.3	81.6	81.3	80.1	84.9	84.8	85.0
4. $q^{\text{img}} \to c^{\text{txt}}$	VisualNews	39.5	32.4	33.3	30.4	40.6	39.4	41.4	41.3
	MSCOCO	91.0	91.8	92.2	90.3	88.5	89.5	88.9	90.1
	Fashion200K	17.2	13.9	14.7	13.2	20.0	17.5	19.9	18.4
5. $q^{\text{img}} \to c^{\text{img}}$	NIGHTS	31.6	31.8	30.7	30.4	31.9	31.8	31.1	32.4
6. $(q^{\text{img}}, q^{\text{txt}}) \to c^{\text{txt}}$	OVEN	40.4	37.9	39.1	36.3	40.9	42.9	42.6	42.1
	InfoSeek	26.1	31.0	32.9	33.3	27.6	37.2	35.8	42.3
7. $(q^{\text{img}}, q^{\text{txt}}) \to c^{\text{img}}$	FashionIQ	24.2	27.4	27.0	26.0	21.7	25.8	26.6	25.7
	CIRR	43.2	48.1	45.4	45.3	38.3	49.5	50.8	50.0
8. $(q^{\text{img}}, q^{\text{txt}}) \to (c^{\text{img}}, c^{\text{txt}})$	OVEN	60.9	61.6	62.6	61.7	61.6	63.9	63.5	64.1
	InfoSeek	45.9	50.3	50.0	53.4	47.1	54.4	53.5	57.7
M-BEIR Avg.	All	47.4	47.9	48.3	47.2	48.3	51.9	52.3	**52.7**
	Single-modal Qry	51.7	51.0	51.6	50.0	53.5	55.6	**56.4**	56.1
	Multi-modal Qry	40.1	42.7	42.8	42.7	39.5	45.6	45.5	**47.0**
MTEB Text Retrieval Avg.		-	-	-	-	-	46.4	49.7	**60.3***

* ranked top-5 on MTEB retrieval task leaderboard. NV-Embed-v1 (Lee et al., 2024) scores 59.36 in MTEB retrieval task.

each multimodal query (or document) input in one word, using embedding for the last token to encode multimodal input;[5] (3) NV-Embed-v1: The LLM from LLaVa-Next is replaced by the fine-tuned LLM-based text retrieval model NV-Embed-v1 (Lee et al., 2024) while all other components (i.e., image encoder and vision–language MLP projector) remain unchanged.[6] Note that the backbone of NV-Embed-v1 is Mistral 7B. The instructions for LLaVa-E (or NV-Embed-v1) and LLaVa-P are illustrated in Table 11 and 12 (in the Appendix), respectively. For reranking experiments, we also utilize LLaVa-Next with Mistral 7B and the prompts are listed in Table 13 (in the Appendix).

Retriever Training Details. For each backbone, we start from fine-tuning M^{rand} with random negatives; i.e., $\mathcal{D}_{\mathcal{B}} = (c_1^+, \cdots, c_{|\mathcal{B}|}^+)$ in Eq. 1. The fine-tuned model is denoted M^{rand}. For CLIP backbone, following (Wei et al., 2023), we fine-tune CLIP$_{\text{SF}}$ for 20 epochs with learning rate $1e-5$. For LLaVa-Next backbone, we fine-tune models for 2 epochs with learning rate $1e-4$. Note that for LLaVa-Next backbone, we only fine-tune the vision–language projector and LoRA ($r = 8, \alpha = 64$) added on the language model. At the stage of fine-tuning M^{hard} with hard negatives, we mine the two types of hard negatives following Section 4.1.1 using each retriever. Then, we fine-tune each retriever using its own mined hard negatives with the same training procedure as the first stage; i.e., $\mathcal{D}_{\mathcal{B}} = (c_1^+, c_1^-, \cdots, c_{|\mathcal{B}|}^+, c_{|\mathcal{B}|}^-)$ in Eq. 1. We fine-tune models with the batch size of 128×8 and 64×8 when using random and hard negatives, respectively. When GPU memory is not enough for the designated batch size, we use gradient accumulation. Note that we initialize retriever using the pre-trained model rather than M^{rand}. We denote the models fine-tuned with random and hard negatives $M^{\text{rand}}(\cdot)$ and $M^{\text{hard}}(\cdot)$, respectively. We refer readers to the Appendix A.1 for more detail.

To enhance text-to-text retrieval capability, we continuously fine-tune M^{hard}(NV-Embed-v1) with learning rate $2e-5$ using the mixture of training data from M-BEIR and public text retrieval datasets aforementioned in Section 4.1.2 for 4.5K steps. The final model is coined MM-Embed.

5.2 Main Results

Universal Multimodal Retrieval. Table 1 reports the retrieval accuracy of different retrievers. In M-BEIR evaluation, we observe that when fine-tuning with random negatives, LLaVa-P achieves the highest overall retrieval effectiveness. This result indicates that LLaVa-P effectively aggregates multimodal input information into a single word representation. While MLLM-based retrievers outperform CLIP$_{\text{SF}}$ on tasks involving multi-modal queries, they still lag behind CLIP$_{\text{SF}}$ on tasks

[5] We refer readers to Table 12 in the Appendix for the prompt and more detail from the prior work (Zhuang et al., 2024; Jiang et al., 2024).

[6] https://huggingface.co/nvidia/NV-Embed-v1

with single-modal queries, especially in cross-modality retrieval; i.e., tasks 1 and 4. In addition, NV-Embed-v1 reaches the best text-to-text retrieval accuracy on WebQA task2.

Observing from the models fine-tuned with hard negatives, MLLM-based retrievers show significant retrieval accuracy improvements, particularly in tasks involving single-modal queries. On the other hand, CLIP_{SF} does not show similar improvement. This could attribute to the fact that CLIP has been well pre-trained for cross-modal retrieval whereas MLLM-based retrievers, fine-tuned with contrastive learning objective for only 2 epochs, may still be underfitting. Fine-tuning with hard negatives accelerates contrastive learning of MLLM-based retrievers.

Table 2: Retrieval analysis on MSCOCO. M.A.@1 denotes the modality accuracy of the top-1 candidate.

Task	Metric	M^{rand}				M^{hard}		
		CLIP_{SF}	LLaVa-E	LLaVa-P	NV-Embed-v1	CLIP_{SF}	LLaVa-P	NV-Embed-v1
1.	R@1	42.6	33.9	41.7	14.1	45.8	50.7	49.8
	R@5	72.0	69.3	70.8	64.6	69.2	73.8	72.7
	M.A.@1	92.6	79.9	91.0	42.1	98.3	100.0	100.0
4.	R@1	72.3	73.0	73.4	69.3	63.8	72.7	72.4
	R@5	91.0	91.8	92.2	90.3	88.5	89.5	88.9
	M.A.@1	98.7	99.2	99.8	96.3	94.2	100.0	100.0

Table 2 reveals another factor contributing to the lower retrieval accuracy of MLLM-based retrievers for single-modal queries: text retrieval bias. This issue is particularly obvious for NV-Embed-v1. We compare models' retrieval accuracy on text–image and image–text retrieval (tasks 1 and 4) on MSCOCO. The comparison shows that M^{rand}(LLaVa-E) and M^{rand}(NV-Embed-v1) exhibit significant lower modality accuracy (M.A.@1) than M^{rand}(CLIP_{SF}) in the text-to-image retrieval task. Most erroneous top-1 retrieved candidates from the MLLM-based retrievers are relevant texts rather than images (see Figure 2 in the Appendix). This result indicates that MLLM-based retrievers have a bias toward relevant text rather than images. This issue can be mitigated by our proposed modality-aware hard negative mining.

Finally, we observe that M^{hard}(NV-Embed-v1) outperforms M^{hard}(LLaVa-P) in text-to-text retrieval tasks (i.e., WebQA task 2 and MTEB); however, compared to the original NV-Embed-v1 (Lee et al., 2024), the score on MTEB retrieval tasks drops almost 10 points. After continual fine-tuning (detailed in Section 4.1.2), the final model, MM-Embed, not only surpasses NV-Embed-v1 in MTEB but also maintains strong multimodal retrieval capability. We attribute the improvement in text-to-text retrieval to the effective hard negatives mined by NV-Embed-v1 aforementioned in Section 4.1.2. Notably, continual fine-tuning significantly enhances multimodal retrieval performance in InfoSeek (col 8 vs 7 in Table 1), highlighting its effectiveness in improving the model's ability to handle knowledge-intensive multimodal retrieval tasks.

Table 3: Experiments of zero-shot reranking on tasks 6–8 from M-BEIR.

Task	Dataset	M^{hard} (NV-Embed-v1)		MM-Embed	
		Ret.	Rerank	Ret.	Rerank
6.	OVEN	42.6	44.3	42.1	43.5
	InfoSeek	35.8	37.1	42.3	43.1
7.	FashionIQ	26.6	20.0	25.7	19.0
	CIRR	50.8	48.6	50.0	48.2
8.	OVEN	63.5	65.8	64.1	65.9
	InfoSeek	53.5	54.5	57.7	57.3

Table 4: Experiments of zero-shot reranking on composed image retrieval task, CIRCO (Baldrati et al., 2023).

Model	Ret.	Rerank
MagicLens (Zhang et al., 2024)	24.9	32.4
E5-V (Jiang et al., 2024)	19.1	31.0
M^{rand}(CLIP_{SF})	12.7	31.6
M^{hard} (LLaVa-P)	29.0	37.9
M^{hard} (NV-Embed-v1)	32.4	40.9
MM-Embed	32.3	39.9

Zero-Shot Reranking. Table 3 reports the reranked results from the top-10 retrieved candidates of M^{hard}(NV-Embed-v1) and MM-Embed on the tasks involving multi-modal queries. We observe accuracy improvements in visual question answering retrieval tasks (i.e., OVEN and InfoSeek), but no improvement on composed image retrieval tasks (i.e., FashionIQ and CIRR). However, as shown in Table 8 (in the Appendix), compared to OVEN and InfoSeek, FashionIQ and CIRR only have one relevance label per query. We hypothesize that there may be additional relevant positives that are not labeled. We refer readers to Figure 3 in the Appendix for case studies.

We conduct experiments on the composed image retrieval dataset with high-quality human annotations, CIRCO (Baldrati et al., 2023) validation set, consisting of 219 queries and 123K candidates in total. On average, 4.2 positives are labeled by humans per query. Table 4 reports mAP@5 for various retrievers and their reranking results. We directly use the models and code provided by the authors to get the results of MagicLens (Zhang et al., 2024)[7] and E5-V (Jiang et al., 2024)[8] retrievers. For our retrievers fine-tuned on M-BEIR, M^{rand}(CLIP$_{\text{SF}}$), M^{hard}(LLaVa-P), M^{hard}(NV-Embed-v1) and MM-Embed, we directly use the same instructions as CIRR in M-BEIR for query encoding. We first observe that our MLLM-based retrievers outperform MagicLens and E5-V. More importantly, reranking upon the top-10 retrieved candidates from the different retrievers significantly improves mAP@5 by at least 7 points. The result demonstrates the effectiveness of prompting an MLLM as a reranker in composed image retrieval tasks.

5.3 Ablation Studies

5.3.1 Is Fine-Tuning with Instruction Necessary?

Table 5: Ablation study on fine-tuning NV-Embed-v1 w/o (✗) and w/ (✓) instructions.

Task	Dataset	zero-shot				fine-tuning	
		CLIP	LLaVa-P	NV-Embed-v1		NV-Embed-v1	
		✗	✗	✗	✓	✗	✓
1. $q^{\text{txt}} \to c^{\text{img}}$	VisualNews	40.9	11.7	15.3	17.4	33.1	38.7
	MSCOCO	55.4	58.1	64.2	59.9	76.7	82.8
	Fashion200K	8.9	2.4	4.2	3.2	12.3	15.6
4. $q^{\text{img}} \to c^{\text{txt}}$	VisualNews	42.0	6.3	6.5	5.9	29.3	37.2
	MSCOCO	79.6	66.8	70.6	68.2	88.9	93.0
	Fashion200K	7.7	2.9	4.0	3.6	12.0	16.8
5. $q^{\text{img}} \to c^{\text{img}}$	NIGHTS	25.4	28.4	29.3	27.7	31.6	30.9

We fine-tune NV-Embed-v1 with random negatives on the M-BEIR subtasks listed in Table 5 and evaluate models' retrieval accuracy on the development queries from each subtask. Note that, for simplicity, we encode only the corpus specific to each dataset, containing documents of the targeted modality. For example, when evaluating retrieval accuracy for VisualNews task 1, we encode the 542K images from VisualNews (see Table 8 in the Appendix) as the index rather than the entire 5.6M documents from M-BEIR. We also report CLIP and LLaVa-P (w/o instruction) zero-shot retrieval effectiveness as a reference point.[9]

From Table 5, we observe that NV-Embed-v1, as a zero-shot MLLM-based retriever, outperforms LLaVa-P and even competes CLIP in the tasks in Miscellaneous domain (i.e., MSCOCO and NIGHTS). This result indicates that a fine-tuned MLLM-based text retriever is capable to perform multimodal retrieval tasks (same finding in (Jiang et al., 2024)). Although incorporating task instructions with queries degrades the retrieval effectiveness (col 4 vs 3), the model fine-tuned with instructions significantly outperforms the one fine-tuned without instructions (col 6 vs 5). This indicates that task instructions can help elicit models' task- or domain-specific knowledge for diverse multimodal retrieval tasks.

5.3.2 Effectiveness of Continual text-to-text retrieval Fine-Tuning

In this section, we study the best strategy to enhance models' capabilities in both multimodal and text-to-text retrieval. We begin by fine-tuning NV-Embed-v1 on both training data for universal multimodal retrieval and text-to-text retrieval (detailed in Section 4.1.2) for 2K steps. As shown in Table 6, joint fine-tuning for both tasks allows the model to maintain its text retrieval capability (row 3 vs 1), although it results in a drop of over 2 points in multimodal retrieval accuracy (row 3 vs 2). In contrast, consciously fine-tuning M^{hard}(NV-Embed-v1) for addition 2K steps significantly boosts its text-to-text retrieval capability with a slight drop of 0.8 points in multimodal retrieval (row 5 vs 4).[10]
This experiment shows that continuously fine-tuning a multimodal retriever to enhance its text-to-text retrieval is more effective than fine-tuning a retriever on all the retrieval tasks simultaneously. This finding suggests that a more optimized curriculum learning strategy (Bengio et al., 2009) could further improve performance in universal multimodal retrieval, a direction we leave for future work.

[7] `https://github.com/google-deepmind/magiclens`

[8] `https://github.com/kongds/E5-V`

[9] We follow Jiang et al. (2024) to prompt LLaVa-Next to output one word embedding for each query and document. i.e., <txt>\nSummary above sentence in one word:; <img>\nSummary above image in one word:.

[10] Note that MM-Embed in Table 1 is fine-tuned with the same condition with total 4.5K steps.

Table 6: Ablation study to enhance model's text-to-text retrieval capability.

Initialization	Training data		M-BEIR*	BEIR*
	Multimodal	Text-to-Text		
NV-Embed-v1	-	-	-	62.9
	✓	✗	54.3	51.7
	✓	✓	52.2	63.0
M^{hard} (NV-Embed-v1)	-	-	56.4	51.7
	✓	✓	55.6	63.1

* For M-BEIR, we evaluate on the tasks with single-modality queries (i.e., tasks 1–5) while for BIER, we evaluate on 7 tasks: ArguAna, FiQA, NFCorpus, Quora, SCIDOCS, SciFact and TREC-COVID.

5.3.3 Study on Prompting MLLMs for Reranking

In this section, we study the reranking effectiveness of MLLMs on all the tasks in M-BEIR dataset. Specifically, for each development query, we rerank the top-10 retrieved candidates from M^{rand}(CLIP$_{\text{SF}}$). As shown in Table 7, prompting LLaVa-Next for reranking further boosts the ranking accuracy in tasks 6–8, which involve multimodal queries (except for FashionIQ). However, the reranking degrades accuracy in tasks 1–5 which involve single-modal queries (except for WebQA task 2). This trend persists even after scaling the reranker from 7B to 34B (col 3, 2 vs 1).[11] We hypothesize that it is challenging for bi-encoder models to encode multimodal queries, such as visual question answering and composed image retrieval. Prompting an MLLM as a reranker in a zero-shot or few-shot manner, or distilling the reranked results into a bi-encoder retriever is a promising solution.

Table 7: Reranking study on top-10 retrieved candidates from M^{rand}(CLIP$_{\text{SF}}$) on M-BEIR development query set.

Task	**Dataset**	Ret.	Rerank	
			7B	34B
1. $q^{\text{txt}} \to c^{\text{img}}$	VisualNews	44.2	38.8	42.5
	MSCOCO	72.0	68.0	69.7
	Fashion200K	17.8	14.7	15.6
2. $q^{\text{txt}} \to c^{\text{txt}}$	WebQA	78.2	79.2	82.9
3. $q^{\text{txt}} \to (c^{\text{img}}, c^{\text{txt}})$	EDIS	48.3	46.5	47.4
	WebQA	78.2	67.7	68.3
4. $q^{\text{img}} \to c^{\text{txt}}$	VisualNews	37.4	29.3	29.8
	MSCOCO	91.0	87.3	89.0
	Fashion200K	17.3	9.9	12.0
5. $q^{\text{img}} \to c^{\text{img}}$	NIGHTS	32.1	29.4	32.7
6. $(q^{\text{img}}, q^{\text{txt}}) \to c^{\text{txt}}$	OVEN	40.6	43.2	43.7
	InfoSeek	25.6	28.4	29.0
7. $(q^{\text{img}}, q^{\text{txt}}) \to c^{\text{img}}$	FashionIQ	32.5	21.5	23.4
	CIRR	52.4	54.1	54.2
8. $(q^{\text{img}}, q^{\text{txt}}) \to (c^{\text{img}}, c^{\text{txt}})$	OVEN	60.6	63.8	63.7
	InfoSeek	45.3	48.7	50.5

6 Conclusion and Future Work

In this paper, we present techniques for advancing information retrieval with multimodal large language models (MLLMs). We first study fine-tuning MLLM-based retrievers to tackle a general information retrieval scenario: universal multimodal retrieval, where models are required to deal with diverse retrieval tasks, multimodal queries and documents. Our study shows that MLLM-based retrievers exhibit *modality bias* in cross-modal retrieval tasks compared to CLIP-based retrievers. To address the issue, we propose modality-aware hard negative mining, which significantly improves our MLLM-based retrievers' accuracy by 5 points in M-BEIR dataset, a benchmark for universal multimodal retrieval. Additionally, with our proposed continual fine-tuning, our MLLM-based retriever, MM-Embed, is the first model to yield state-of-the-art retrieval accuracy in universal multimodal retrieval tasks while maintaining strong text-to-text retrieval capability (ranked top-5 on MTEB retrieval task leaderboard). Finally, we explore to prompt MLLMs as reranker in M-BEIR tasks. We find that MLLMs can be used as zero-shot rerankers to further boost retrieval accuracy in the challenging tasks, which require the understanding of multimodal queries, such as visual question answering and composed image retrieval. For example, our zero-shot MLLM-based reranker improves the retrieval accuracy upon the state-of-the-art retrievers by over 7 points in CIRCO.

[11] In the experiment, we use the model from https://huggingface.co/llava-hf/llava-v1.6-34b-hf

Our work also suggests two promising future directions: (1) Distilling our MLLM-based retriever, MM-Embed, to smaller multimodal retrievers, such as CLIP (Radford et al., 2021) or BLIP (Li et al., 2022); (2) Distilling MLLM-based reranker to retriever to further improve its retrieval capability in tasks involving multimodal queries. In addition, recent work (Ma et al., 2024a; Faysse et al., 2024) has demonstrated that MLLMs can be fine-tuned to tackle visual document retrieval tasks, which could be integrated into universal multimodal retrieval.

A APPENDIX

Table 8: M-BEIR dataset statistics.

Task	Dataset	Domain	# Query			# Relevance / Query			# Candid.
			Train	Dev	Test	Train	Dev	Test	
1. $q^{\text{txt}} \to c^{\text{img}}$	VisualNews (Liu et al., 2021a)	News	99K	20K	20K	1.0	1.0	1.0	542K
	MSCOCO (Lin et al., 2014)	Misc.	100K	24.8K	24.8K	1.0	1.0	1.0	5K
	Fashion200K (Han et al., 2017)	Fashion	15K	1.7K	1.7K	3.3	3.1	2.8	201K
2. $q^{\text{txt}} \to c^{\text{txt}}$	WebQA (Chang et al., 2022)	Wiki	16K	1.7K	2.4K	2.0	2.0	2.0	544K
3. $q^{\text{txt}} \to (c^{\text{img}}, c^{\text{txt}})$	EDIS (Liu et al., 2023b)	News	26K	3.2K	3.2K	2.6	2.6	2.6	1M
	WebQA (Chang et al., 2022)	Wiki	16K	1.7K	2.4K	1.4	1.4	1.4	544K
4. $q^{\text{img}} \to c^{\text{txt}}$	VisualNews (Liu et al., 2021a)	News	100K	20K	20K	1.0	1.0	1.0	537K
	MSCOCO (Lin et al., 2014)	Misc.	113K	5K	5K	5.0	5.0	5.0	25K
	Fashion200K (Han et al., 2017)	Fashion	15K	4.8K	4.8K	1.0	1.0	1.0	61K
5. $q^{\text{img}} \to c^{\text{img}}$	NIGHTS (Fu et al., 2023)	Misc.	16K	2K	2K	1.0	1.0	1.0	40K
6. $(q^{\text{img}}, q^{\text{txt}}) \to c^{\text{txt}}$	OVEN (Hu et al., 2023)	Wiki	150K	50K	50K	8.5	10.0	9.9	676K
	InfoSeek (Chen et al., 2023)	Wiki	141K	11K	11K	6.8	6.7	6.5	611K
7. $(q^{\text{img}}, q^{\text{txt}}) \to c^{\text{img}}$	FashionIQ (Wu et al., 2021)	Fashion	16K	2K	6K	1.0	1.0	1.0	74K
	CIRR (Liu et al., 2021b)	Misc.	26K	2K	4K	1.0	1.0	1.0	21K
8. $(q^{\text{img}}, q^{\text{txt}}) \to (c^{\text{img}}, c^{\text{txt}})$	OVEN (Hu et al., 2023)	Wiki	157K	14.7K	14.7K	17.8	17.5	17.7	335K
	InfoSeek (Chen et al., 2023)	Wiki	143K	17.6K	17.6K	9.1	7.5	7.5	481K
M-BEIR (Wei et al., 2023)		4 domains	1.1M	182K	190K	6.5	5.9	5.7	5.6M

Table 9: Detailed results on MTEB retrieval tasks.

Model	AA	CF	CQ	DB	Fe	FQ	HQ	MS	NF	NQ	Qu	SD	SF	T2	TC	Avg.
NV-Embed-v1 (Lee et al., 2024)	68.2	34.7	50.5	48.3	87.8	63.1	79.9	46.5	38.0	71.2	89.2	20.2	78.4	28.4	85.9	59.4
M^{hard} (LLaVa-P)	38.6	20.4	38.0	36.9	78.1	36.2	61.2	23.2	35.1	45.1	86.1	19.2	72.7	27.7	77.2	46.4
M^{hard} (NV-Embed-v1)	37.2	30.8	44.0	44.3	86.4	45.5	70.6	34.2	37.4	49.7	86.9	13.9	64.1	23.5	76.7	49.7
MM-Embed	69.0	39.3	49.7	50.6	92.6	60.1	81.4	45.1	40.5	70.6	88.7	21.8	78.3	31.1	85.4	60.3

* Dataset Legend: AA=ArguAna, CF=Climate-FEVER, CQ=CQADupStack, DB=DBPedia, Fe=FEVER, FQ=FiQA, HQ=HotpotQA, MS=MSMARCO, NF=NFCorpus, NQ=Natural Questions, Qu=Quora, SD=SCIDOCS, SF=SciFact, T2=Touché-2020, TC=TREC-COVID

A.1 IMPLEMENTATION DETAILS

We implement our training and inference using Tevatron (Gao et al., 2023). For CLIP-based retrievers, we follow all the settings from Wei et al. (2023). For MLLM-based retriever, we fine-tune models with DeepSpeed Zero 2 (Rajbhandari et al., 2020) and gradient checkpointing. During fine-tuning on M-BEIR training data, we set maximum length for queries and documents to 128. While continual fine-tuning on both M-BEIR and text-to-text retrieval training data, we set maximum length for queries and documents to 128 and 512, respectively. All fine-tuning are conducted on 8×80GB A100 GPUs. Note that image input only occupies single token length after being tokenized; however, each image will be converted to multiple image tokens. Thus, the actual input length to MLLM is longer than the maximum length we set. To speed fine-tuning and inference for MLLM-based retrievers, we only use the global image patches, which occupy 576 (24×24) image tokens.

A.2 BASELINE REPRODUCING

Since we implement our fine-tuning and inference following the setting from Wei et al. (2023), our fine-tuned M^{rand}(CLIP$_{\text{SF}}$) should be equal to CLIP$_{\text{SF}}$from Wei et al. (2023). In Table 10, we compare the results from our fine-tuned M^{rand}(CLIP$_{\text{SF}}$) and the checkpoint provided by the authors.[12]

Table 10: A comparison of M^{rand}(CLIP$_{\text{SF}}$) fine-tuned by us and Wei et al. (2023).

Task	Dataset	M^{rand}(CLIP$_{\text{SF}}$)	
		Wei et al. (2023)	Ours
M-BEIR Avg.	All	47.4	47.4
	Single-modal Qry	52.5	51.7
	multi-modal Qry	39.1	40.1

[12] https://huggingface.co/TIGER-Lab/UniIR/blob/main/checkpoint/CLIP_SF/clip_sf_large.pth

Table 11: NV-Embed-v1 (and LLaVa-E) instructions for M-BEIR and MTEB, which are from Wei et al. (2023) and Lee et al. (2024), respectively. For all the candidates, we use the prompt to generate the embedding: $< c^{img} >$\n$< c^{txt} >$<eos>.

Task	Dataset	M-BEIR task instruction
1. $q^{\text{txt}} \to c^{\text{img}}$	VisualNews	*Identify the news-related image in line with the described event.\nQuery:* $< q^{txt} >$<eos>
	MSCOCO	*Find me an everyday image that matches the given caption.\nQuery:* $< q^{txt} >$<eos>
	Fashion200K	*Based on the following fashion description, retrieve the best matching image.\nQuery:* $< q^{txt} >$<eos>
2. $q^{\text{txt}} \to c^{\text{txt}}$	WebQA	*Retrieve passages from Wikipedia that provide answers to the following question.\nQuery:* $< q^{txt} >$<eos>
3. $q^{\text{txt}} \to (c^{\text{img}}, c^{\text{txt}})$	EDIS	*Find a news image that matches the provided caption.\nQuery:* $< q^{txt} >$<eos>
	WebQA	*Find a Wikipedia image that answers this question.\nQuery:* $< q^{txt} >$<eos>
4. $q^{\text{img}} \to c^{\text{txt}}$	VisualNews	*Find a caption for the news in the given photo.\nQuery:* $< q^{img} >$<eos>
	MSCOCO	*Find an image caption describing the following everyday image.\nQuery:* $< q^{img} >$<eos>
	Fashion200K	*Find a product description for the fashion item in the image.\nQuery:* $< q^{img} >$<eos>
5. $q^{\text{img}} \to c^{\text{img}}$	NIGHTS	*Find a day-to-day image that looks similar to the provided image.\nQuery:* $< q^{img} >$<eos>
6. $(q^{\text{img}}, q^{\text{txt}}) \to c^{\text{txt}}$	OVEN	*Retrieve a Wikipedia paragraph that provides an answer to the given query about the image.\nQuery:* $< q^{img} >$\n$< q^{img} >$<eos>
	InfoSeek	*Retrieve a Wikipedia paragraph that provides an answer to the given query about the image.\nQuery:* $< q^{img} >$\n$< q^{img} >$<eos>
7. $(q^{\text{img}}, q^{\text{txt}}) \to c^{\text{img}}$	FashionIQ	*Find a fashion image that aligns with the reference image and style note.\nQuery:* $< q^{img} >$\n$< q^{img} >$<eos>
	CIRR	*Retrieve a day-to-day image that aligns with the modification instructions of the provided image.\nQuery:* $< q^{img} >$\n$< q^{img} >$<eos>
8. $(q^{\text{img}}, q^{\text{txt}}) \to (c^{\text{img}}, c^{\text{txt}})$	OVEN	*Retrieve a Wikipedia image-description pair that provides evidence for the question of this image.\nQuery:* $< q^{img} >$\n$< q^{img} >$<eos>
	InfoSeek	*Retrieve a Wikipedia image-description pair that provides evidence for the question of this image.\nQuery:* $< q^{img} >$\n$< q^{img} >$<eos>
Task	**Dataset**	**MTEB task instruction**
9. $q^{\text{txt}} \to c^{\text{txt}}$	ArguAna	*Given a claim, find documents that refute the claim\nQuery:* $< q^{txt} >$<eos>
	Climate-FEVER	*Given a claim about climate change, retrieve documents that support or refute the claim\nQuery:* $< q^{txt} >$<eos>
	CQADupStack	*Given a question, retrieve detailed question descriptions from StackExchange that are duplicates to the given question\nQuery:* $< q^{txt} >$<eos>
	DBPedia	*Given a query, retrieve relevant entity descriptions from DBPedia\nQuery:* $< q^{txt} >$<eos>
	FEVER	Given a claim, retrieve documents that support or refute the claim\nQuery: $< q^{txt} >$<eos>
	FiQA	*Given a financial question, retrieve user replies that best answer the question\nQuery:* $< q^{txt} >$<eos>
	HotpotQA	*Given a multi-hop question, retrieve documents that can help answer the question\nQuery:* $< q^{txt} >$<eos>
	MSMARCO	*Given a web search query, retrieve relevant passages that answer the query\nQuery:* $< q^{txt} >$<eos>
	NFCorpus	*Given a question, retrieve relevant documents that best answer the question\nQuery:* $< q^{txt} >$<eos>
	Natural Questions	*Given a question, retrieve Wikipedia passages that answer the question\nQuery:* $< q^{txt} >$<eos>
	Quora	*Find questions that have the same meaning as the input question\nQuery:* $< q^{txt} >$<eos>
	SCIDOCS	*Given a scientific paper title, retrieve paper abstracts that are cited by the given paper\nQuery:* $< q^{txt} >$<eos>
	SciFact	*Given a scientific claim, retrieve documents that support or refute the claim\nQuery:* $< q^{txt} >$<eos>
	Touch´e-2020	*Given a question, retrieve detailed and persuasive arguments that answer the question\nQuery:* $< q^{txt} >$<eos>
	TREC-COVID	*Given a query on COVID-19, retrieve documents that answer the query\nQuery:* $< q^{txt} >$<eos>

Table 12: LLaVa-P instructions for M-BEIR and MTEB. [image], [text] and [image,text] are used to inform LLaVa-P the user desired modality. For all the candidates, we use the prompt to generate the embedding: $< c^{img} >$\n$< c^{txt} >$*\nDescribe the above in one word:*

Task	Dataset	M-BEIR task instruction
1. $q^{\text{txt}} \to c^{\text{img}}$	VisualNews	*[image]* $< q^{txt} >$*\nDescribe the news-related caption in one word:*
	MSCOCO	*[image]* $< q^{txt} >$*\nDescribe the everyday caption in one word:*
	Fashion200K	*[image]* $< q^{txt} >$*\nDescribe the fashion description in one word:*
2. $q^{\text{txt}} \to c^{\text{txt}}$	WebQA	*[text]* $< q^{txt} >$*\nAnswer the question using Wikipedia in one word:*
3. $q^{\text{txt}} \to (c^{\text{img}}, c^{\text{txt}})$	EDIS	*[image,text]* $< q^{txt} >$*\nDescribe the news-related caption in one word:*
	WebQA	*[image,text]* $< q^{txt} >$*\nAnswer the question using Wikipedia in one word:*
4. $q^{\text{img}} \to c^{\text{txt}}$	VisualNews	*[text]* $< q^{img} >$*\nDescribe the news-related image in one word:*
	MSCOCO	*[text]* $< q^{img} >$*\nDescribe the everyday image in one word:*
	Fashion200K	*[text]* $< q^{img} >$*\nDescribe the fashion image in one word:*
5. $q^{\text{img}} \to c^{\text{img}}$	NIGHTS	*[image]* $< q^{img} >$*\nDescribe the everyday image in one word:*
6. $(q^{\text{img}}, q^{\text{txt}}) \to c^{\text{txt}}$	OVEN InfoSeek	[text] $< q^{img} >$\n$< q^{txt} >$*\nAnswer the question based on the image from Wikipedia in one word:*
7. $(q^{\text{img}}, q^{\text{txt}}) \to c^{\text{img}}$	FashionIQ	*[image]* $< q^{img} >$*\nChange the style of this shirt/dress/toptee to* $< q^{txt} >$*\nDescribe this modified shirt/dress/toptee in one word:*
	CIRR	*[image]* $< q^{img} >$*\nModify this image with* $< q^{txt} >$*\nDesribe modified image in one word:*
8. $(q^{\text{img}}, q^{\text{txt}}) \to (c^{\text{img}}, c^{\text{txt}})$	OVEN InfoSeek	*[image,text]* $< q^{img} >$\n$< q^{txt} >$*\nAnswer the question based on the interleaved image-text passage from Wikipedia in one word:*
Task	**Dataset**	**MTEB task instruction**
9. $q^{\text{txt}} \to c^{\text{txt}}$	ArguAna	*[text]* $< q^{txt} >$*\nGiven a claim, generate a document that refute the claim in one word:*
	Climate-FEVER	*[text]* $< q^{txt} >$*\nGiven a claim about climate change, generate a document that supports or refutes the claim in one word:*
	CQADupStack	*[text]* $< q^{txt} >$*\nDescribe the StackExchange question in one word:*
	DBPedia	*[text]* $< q^{txt} >$*\nGiven a query, generate a relevant entity description from DBPedia in one word:*
	FEVER	*[text]* $< q^{txt} >$*\nGiven a claim, generate a document that supports or refutes the claim in one word:*
	FiQA	*[text]* $< q^{txt} >$*\nAnswer the financial question in one word:*
	HotpotQA	*[text]* $< q^{txt} >$*\nAnswer the multi-hop question in one word:*
	MSMARCO	*[text]* $< q^{txt} >$*\nAnswer the web search query in one word:*
	NFCorpus	*[text]* $< q^{txt} >$*\nAnswer the question in one word:*
	Natural Questions	*[text]* $< q^{txt} >$*\nAnswer the question using Wikipedia in one word:*
	Quora	*[text]* $< q^{txt} >$*\nDescribe the question in one word:*
	SCIDOCS	*[text]* $< q^{txt} >$*\nGiven a scientific paper title, generate a paper abstract that is cited by the given paper in one word:*
	SciFact	*[text]* $< q^{txt} >$*\nGiven a scientific claim, generate a document that support or refute the claim in one word:*
	Touch´e-2020	*[text]* $< q^{txt} >$*\nAnswer the question with detailed and persuasive arguments in one word:*
	TREC-COVID	*[text]* $< q^{txt} >$*\nAnswer the query on COVID-19 in one word:*

Table 13: Prompts for reranking tasks in M-BEIR .

Task	Dataset	Prompt
1. $q^{\text{txt}} \to c^{\text{img}}$	VisualNews	$< c^{img} >$*\nNews:*$< q^{txt} >$*\nDoes the above News image match the News story? True or False*
	MSCOCO	$< c^{img} >$*\nCaption:*$< q^{txt} >$*\nDoes the above daily-life image match the caption? True or False*
	Fashion200K	$< c^{img} >$*\nDescription:*$< q^{txt} >$*\nDoes the above image match the cloth style description? True or False*
2. $q^{\text{txt}} \to c^{\text{txt}}$	WebQA	*Question:* $< q^{txt} >$*\nAnswer:* $< c^{txt} >$*\nDoes the answer correctly answer the question? True or False*
3. $q^{\text{txt}} \to (c^{\text{img}}, c^{\text{txt}})$	EDIS WebQA	*Question:* $< q^{txt} >$*\nAnswer:* $< c^{txt} >$*\nDoes the answer correctly answer the question? True or False*
4. $q^{\text{img}} \to c^{\text{txt}}$	VisualNews	$< q^{img} >$*\nNews:*$< c^{txt} >$*\nDoes the above News image match the News story? True or False*
	MSCOCO	$< q^{img} >$*\nCaption:*$< c^{txt} >$*\nDoes the above daily-life image match the caption? True or False*
	Fashion200K	$< q^{img} >$*\nDescription:*$< c^{txt} >$*\nDoes the above image match the cloth style description? True or False*
5. $q^{\text{img}} \to c^{\text{img}}$	NIGHTS	$< q^{img} >$\n$< c^{img} >$*\nDoes the above two images have the same scene? True or False*
6. $(q^{\text{img}}, q^{\text{txt}}) \to c^{\text{txt}}$	OVEN InfoSeek	$< q^{img} >$*\nQuestion:*$< q^{txt} >$*\nAnswer:*$< c^{txt} >$*Does the answer correctly answer the question? True or False*
7. $(q^{\text{img}}, q^{\text{txt}}) \to c^{\text{img}}$	FashionIQ CIRR	$< c^{img} >$*\nCaption:*$< q^{txt} >$*\nDoes the above caption describe the modification of the image? True or False*
8. $(q^{\text{img}}, q^{\text{txt}}) \to (c^{\text{img}}, c^{\text{txt}})$	OVEN InfoSeek	$< q^{img} >$*\nQuestion:*$< q^{txt} >$*\nAnswer:*$< c^{txt} >$*Does the answer correctly answer the question? True or False*

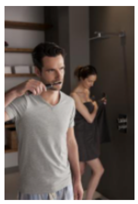
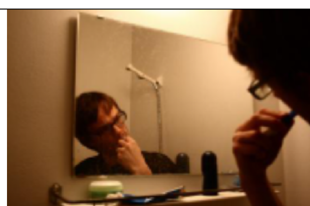
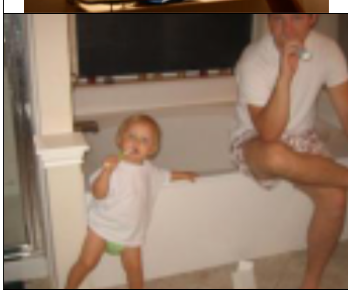
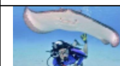
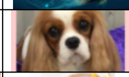
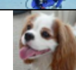
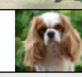
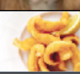
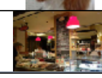
Instruction: Find me an everyday image that matches the given caption. **Query**: A man brushes his teeth while a woman wraps in a towel.	**Instruction**: Identify the news-related image in line with the described event. **Query**: The Q Street NW entrance to the Dupont Circle Metro station.	**Instruction**: Find me an everyday image that matches the given caption. **Query**: A large tow truck towing a double decker bus.
Correct Answer		
		
Negative samples with incorrect modality		
A man brushes his teeth while a woman behind him wraps a towel around herself.	Dupont Circle metro station, Q Street escalator.	A tow truck towing a double decker bus.
A man brushing his teeth with woman in wrapping herself in a towel in the background.	Riding escalator to Q Street exit of Dupont Circle Metro.	A tow truck is in front of a double decker bus.
Negative samples with unsatisfactory information needs		
		
		

Figure 2: Examples of modality-aware negative samples mined by M^{rand}(NV-Embed-v1). We observe that negative samples with incorrect modality show similar semantic meaning as queries while negative samples with unsatisfactory information needs show less accurate information compared to the correct answers

M-BEIR CIRR Task 7				
Query		Answer	Retrieval	Reranking
Human and one animal from a different specie				
Same breed dog, focus on its head.				
Put the fries in a white plate with white background, clean.				

M-BEIR FashionIQ Task 7				
Query		Answer	Retrieval	Reranking
Is shiny and silver with shorter sleeves and fit and flare.				
Is grey with black design and is a light printed short dress.				
Is a solid red color and shorter and tighter with more blue and white.				

Figure 3: Top-1 candidates for the tasks of composed image retrieval and reranking. In many cases, retrieval and reranking yields different top-1 results from labeled positives but appears to be correct since each query only has single labeled positive candidate (see Table 8).

REFERENCES

Akari Asai, Timo Schick, Patrick Lewis, Xilun Chen, Gautier Izacard, Sebastian Riedel, Hannaneh Hajishirzi, and Wen-tau Yih. Task-aware retrieval with instructions. In *Findings of ACL*, pp. 3650–3675, 2023.

Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, et al. MS MARCO: A human generated machine reading comprehension dataset. *arXiv:1611.09268*, 2016.

Alberto Baldrati, Lorenzo Agnolucci, Marco Bertini, and Alberto Del Bimbo. Zero-shot composed image retrieval with textual inversion. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 15338–15347, 2023.

Yoshua Bengio, Jérôme Louradour, Ronan Collobert, and Jason Weston. Curriculum learning. In *Proc. ICML*, pp. 41–48, 2009.

Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. A large annotated corpus for learning natural language inference. In *Proc. EMNLP*, pp. 632–642, 2015.

Yingshan Chang, Guihong Cao, Mridu Narang, Jianfeng Gao, Hisami Suzuki, and Yonatan Bisk. WebQA: Multihop and multimodal qa. In *Proc. CVPR*, pp. 16474–16483, 2022.

Xilun Chen, Kushal Lakhotia, Barlas Oguz, Anchit Gupta, Patrick Lewis, Stan Peshterliev, Yashar Mehdad, Sonal Gupta, and Wen-tau Yih. Salient phrase aware dense retrieval: Can a dense retriever imitate a sparse one? In *Proc. Findings of EMNLP*, pp. 250–262, 2022.

Yang Chen, Hexiang Hu, Yi Luan, Haitian Sun, Soravit Changpinyo, Alan Ritter, and Ming-Wei Chang. Can pre-trained vision and language models answer visual information-seeking questions? In *Proc. EMNLP*, pp. 14948–14968, 2023.

Wenliang Dai, Nayeon Lee, Boxin Wang, Zhuoling Yang, Zihan Liu, Jon Barker, Tuomas Rintamaki, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. NVLM: Open frontier-class multimodal LLMs. *arXiv:2409.11402*, 2024.

Gabriel de Souza P. Moreira, Radek Osmulski, Mengyao Xu, Ronay Ak, Benedikt Schifferer, and Even Oldridge. NV-Retriever: Improving text embedding models with effective hard-negative mining. *arxiv.2407.15831*, 2024.

Manuel Faysse, Hugues Sibille, Tony Wu, Bilel Omrani, Gautier Viaud, Céline Hudelot, and Pierre Colombo. ColPali: Efficient document retrieval with vision language models. *arXiv:2407.01449*, 2024.

Stephanie Fu, Netanel Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. DreamSim: Learning new dimensions of human visual similarity using synthetic data. In *Proc. NeurIPS*, pp. 50742–50768, 2023.

Luyu Gao, Xueguang Ma, Jimmy Lin, and Jamie Callan. Tevatron: An efficient and flexible toolkit for neural retrieval. In *Proc. SIGIR*, pp. 3120–3124, 2023.

Michael Gutmann and Aapo Hyvärinen. Noise-contrastive estimation: A new estimation principle for unnormalized statistical models. In *Proc. AISTATS*, pp. 297–304, 2010.

X. Han, Z. Wu, P. X. Huang, X. Zhang, M. Zhu, Y. Li, Y. Zhao, and L. S. Davis. Automatic spatially-aware fashion concept discovery. In *Proc. ICCV*, pp. 1472–1480, 2017.

Faegheh Hasibi, Fedor Nikolaev, Chenyan Xiong, Krisztian Balog, Svein Erik Bratsberg, Alexander Kotov, and Jamie Callan. DBpedia-Entity v2: A test collection for entity search. In *Proc. SIGIR*, pp. 1265–1268, 2017.

Hexiang Hu, Yi Luan, Yang Chen, Urvashi Khandelwal, Mandar Joshi, Kenton Lee, Kristina Toutanova, and Ming-Wei Chang. Open-domain Visual Entity Recognition: Towards recognizing millions of wikipedia entities. In *Proc. ICCV*, 2023.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. Mistral 7B. *arXiv:2310.06825*, 2023.

Ting Jiang, Minghui Song, Zihan Zhang, Haizhen Huang, Weiwei Deng, Feng Sun, Qi Zhang, Deqing Wang, and Fuzhen Zhuang. E5-V: Universal embeddings with multimodal large language models. *arxiv.2407.12580*, 2024.

Can Jin, Hongwu Peng, Shiyu Zhao, Zhenting Wang, Wujiang Xu, Ligong Han, Jiahui Zhao, Kai Zhong, Sanguthevar Rajasekaran, and Dimitris N. Metaxas. APEER: Automatic prompt engineering enhances large language model reranking. *arXiv:2406.14449*, 2024.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, pp. 535–547, 2021.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering. In *Proc. EMNLP*, pp. 6769–6781, 2020.

Andreas Koukounas, Georgios Mastrapas, Michael Günther, Bo Wang, Scott Martens, Isabelle Mohr, Saba Sturua, Mohammad Kalim Akram, Joan Fontanals Martínez, Saahil Ognawala, Susana Guzman, Maximilian Werk, Nan Wang, and Han Xiao. Jina CLIP: Your CLIP model is also your text retriever. *arXiv:2405.20204*, 2024.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. Natural Questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, pp. 452–466, 2019.

Chankyu Lee, Rajarshi Roy, Mengyao Xu, Jonathan Raiman, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. NV-Embed: Improved techniques for training LLMs as generalist embedding models. *arXiv:2405.17428*, 2024.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Proc. NeurIPS*, pp. 9459–9474, 2020.

Patrick Lewis, Yuxiang Wu, Linqing Liu, Pasquale Minervini, Heinrich Küttler, Aleksandra Piktus, Pontus Stenetorp, and Sebastian Riedel. PAQ: 65 million probably-asked questions and what you can do with them. *Transactions of the Association for Computational Linguistics*, pp. 1098–1115, 2021.

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *Proceedings of the 39th International Conference on Machine Learning*, Proc. ICML, pp. 12888–12900, 2022.

Jimmy Lin, Rodrigo Nogueira, and Andrew Yates. *Pretrained Transformers for Text Ranking: BERT and Beyond*. Morgan & Claypool, 2021.

Sheng-Chieh Lin, Akari Asai, Minghan Li, Barlas Oguz, Jimmy Lin, Yashar Mehdad, Wen-tau Yih, and Xilun Chen. How to Train your DRAGON: Diverse augmentation towards generalizable dense retrieval. In *Proc. Findings of EMNLP*, pp. 6385–6400, 2023.

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: Common objects in context. In *Proc. ECCV*, pp. 740–755, Cham, 2014.

Fuxiao Liu, Yinghan Wang, Tianlu Wang, and Vicente Ordonez. Visual News: Benchmark and challenges in news image captioning. In *Proc. EMNLP*, pp. 6761–6771, 2021a.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Proc. NeurIPS*, volume 36, pp. 34892–34916, 2023a.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. LLaVA-NeXT: Improved reasoning, OCR, and world knowledge, January 2024. URL https://llava-vl.github.io/blog/2024-01-30-llava-next/.

Siqi Liu, Weixi Feng, Tsu-Jui Fu, Wenhu Chen, and William Wang. EDIS: Entity-driven image search over multimodal web content. In *Proc. EMNLP*, pp. 4877–4894, 2023b.

Zhenghao Liu, Chenyan Xiong, Yuanhuiyi Lv, Zhiyuan Liu, and Ge Yu. Universal vision-language dense retrieval: Learning a unified representation space for multi-modal retrieval. In *Proc. ICLR*, 2023c.

Zheyuan Liu, Cristian Rodriguez-Opazo, Damien Teney, and Stephen Gould. Image retrieval on real-life images with pre-trained vision-and-language models. In *Proc. ICCV*, pp. 2105–2114, 2021b.

Xueguang Ma, Sheng-Chieh Lin, Minghan Li, Wenhu Chen, and Jimmy Lin. Unifying multimodal retrieval via document screenshot embedding, 2024a.

Xueguang Ma, Liang Wang, Nan Yang, Furu Wei, and Jimmy Lin. Fine-tuning LLaMA for multi-stage text retrieval. In *Proc. SIGIR*, pp. 2421–2425, 2024b.

Macedo Maia, Siegfried Handschuh, André Freitas, Brian Davis, Ross McDermott, Manel Zarrouk, and Alexandra Balahur. WWW'18 open challenge: Financial opinion mining and question answering. In *Companion Proceedings of the The Web Conference 2018*, pp. 1941–1942, 2018.

Rui Meng, Ye Liu, Shafiq Rayhan Joty, Caiming Xiong, Yingbo Zhou, and Semih Yavuz. SFR-Embedding-2: Advanced text embedding with multi-stage training, 2024. URL https://huggingface.co/Salesforce/SFR-Embedding-2_R.

Gabriel de Souza P Moreira, Radek Osmulski, Mengyao Xu, Ronay Ak, Benedikt Schifferer, and Even Oldridge. Nv-retriever: Improving text embedding models with effective hard-negative mining. *arXiv preprint arXiv:2407.15831*, 2024.

Niklas Muennighoff, Nouamane Tazi, Loic Magne, and Nils Reimers. MTEB: Massive text embedding benchmark. In *Proc. EACL*, pp. 2014–2037, 2023.

Anastasios Nentidis, Georgios Katsimpras, Anastasia Krithara, Salvador Lima López, Eulália Farré-Maduell, Luis Gasco, Martin Krallinger, and Georgios Paliouras. Overview of BioASQ 2023: The eleventh BioASQ challenge on large-scale biomedical semantic indexing and question answering. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pp. 227–250, 2023.

Rodrigo Nogueira, Zhiying Jiang, Ronak Pradeep, and Jimmy Lin. Document ranking with a pre-trained sequence-to-sequence model. In *Proc. Findings of EMNLP*, pp. 708–718, 2020.

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. Learning transferable visual models from natural language supervision. In *Proc. ICML*, pp. 8748–8763, 2021.

Samyam Rajbhandari, Jeff Rasley, Olatunji Ruwase, and Yuxiong He. ZeRO: memory optimizations toward training trillion parameter models. In *Proc. of the International Conference for High Performance Computing, Networking, Storage and Analysis*, 2020.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ questions for machine comprehension of text. In *Proc. EMNLP*, pp. 2383–2392, 2016.

Stack-Exchange-Community. Stack exchange data dump. *Transactions of the Association for Computational Linguistics*, 2023.

Hongjin Su, Weijia Shi, Jungo Kasai, Yizhong Wang, Yushi Hu, Mari Ostendorf, Wen-tau Yih, Noah A. Smith, Luke Zettlemoyer, and Tao Yu. One embedder, any task: Instruction-finetuned text embeddings. In *Proc. Findings of ACL*, pp. 1102–1121, July 2023.

Weiwei Sun, Lingyong Yan, Xinyu Ma, Shuaiqiang Wang, Pengjie Ren, Zhumin Chen, Dawei Yin, and Zhaochun Ren. Is ChatGPT good at search? Investigating large language models as re-ranking agents. In *Proc. EMNLP*, pp. 14918–14937, December 2023.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Proc. NeurIPS*, 2021.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. FEVER: a large-scale dataset for fact extraction and VERification. In *Proc. ACL*, pp. 809–819, 2018.

Henning Wachsmuth, Shahbaz Syed, and Benno Stein. Retrieval of the best counterargument without prior topic knowledge. In *Proc. ACL*, pp. 241–251, 2018.

Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. Improving text embeddings with large language models. *arXiv:2401.00368*, 2023.

Cong Wei, Yang Chen, Haonan Chen, Hexiang Hu, Ge Zhang, Jie Fu, Alan Ritter, and Wenhu Chen. UniIR: Training and benchmarking universal multimodal information retrievers. *arxiv.2311.17136*, 2023.

Hui Wu, Yupeng Gao, Xiaoxiao Guo, Ziad Al-Halah, Steven Rennie, Kristen Grauman, and Rogerio Feris. The Fashion IQ dataset: Retrieving images by combining side information and relative natural language feedback. In *Proc. CVPR*, 2021.

Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul N. Bennett, Junaid Ahmed, and Arnold Overwijk. Approximate nearest neighbor negative contrastive learning for dense text retrieval. In *Proc. ICLR*, 2021.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proc. EMNLP*, pp. 2369–2380, 2018.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, pp. 67–78, 2014.

Kai Zhang, Yi Luan, Hexiang Hu, Kenton Lee, Siyuan Qiao, Wenhu Chen, Yu Su, and Ming-Wei Chang. MagicLens: Self-supervised image retrieval with open-ended instructions. In *Proc. ICML*, pp. to appear, 2024.

Shengyao Zhuang, Xueguang Ma, Bevan Koopman, Jimmy Lin, and Guido Zuccon. PromptReps: Prompting large language models to generate dense and sparse representations for zero-shot document retrieval. *arXiv:2404.18424*, 2024.