Bridging Modalities: Improving Universal Multimodal Retrieval by Multimodal Large Language Models

Xin Zhang[1]*, Yanzhao Zhang[2]*, Wen Xie[2]*, Mingxin Li[2], Ziqi Dai[2], Dingkun Long[2]
Pengjun Xie[2], Meishan Zhang†, Wenjie Li[1], Min Zhang[3]
[1]The Hong Kong Polytechnic University [2]Tongyi Lab, Alibaba Group [3]Soochow University
xin404.zhang@connect.polyu.hk {zhangyanzhao.zyz,dingkun.ldk}@alibaba-inc.com

Abstract

Universal Multimodal Retrieval (UMR) aims to enable search across various modalities using a unified model, where queries and candidates can consist of pure text, images, or a combination of both. Previous work has attempted to adopt multimodal large language models (MLLMs) to realize UMR using only text data. However, our preliminary experiments demonstrate that more diverse multimodal training data can further unlock the potential of MLLMs. Despite its effectiveness, the existing multimodal training data is highly imbalanced in terms of modality, which motivates us to develop a training data synthesis pipeline and construct a large-scale, high-quality fused-modal training dataset. Based on the synthetic training data, we develop the General Multimodal Embedder (GME), an MLLM-based dense retriever designed for UMR. Furthermore, we construct a comprehensive UMR Benchmark (UMRB) to evaluate the effectiveness of our approach. Experimental results show that our method achieves state-of-the-art performance among existing UMR methods. Last, we provide in-depth analyses of model scaling and training strategies, and perform ablation studies on both the model and synthetic data.

1. Introduction

The growth of multimedia applications necessitates retrieval models that extend beyond traditional text-to-text and text-to-image search [75]. In Universal Multimodal Retrieval (UMR) tasks, both queries and candidates can exist in any modality [39]. Compared to addressing this challenge with separate uni-modal and cross-modal retrievers in a divide-and-conquer pipeline [4], a unified retriever is a more viable option in terms of usability and scalability. Using the dense retrieval paradigm (also known as embedding-based retrieval) [25], a unified model can be trained to project inputs from various modalities into a shared embedding space [22, 74, 75]. In this space, similarity scores are computed between the embeddings of queries and the retrieval collection, facilitating the efficient ranking of the top-k candidates. To achieve this, some previous studies have primarily focused on two approaches: (1) designing feature fusion mechanisms for cross-modal retrievers based on the CLIP architecture [39, 66], and (2) incorporating visual plugin modules into optimized text embedding models to achieve unified multimodal representations [74, 75].

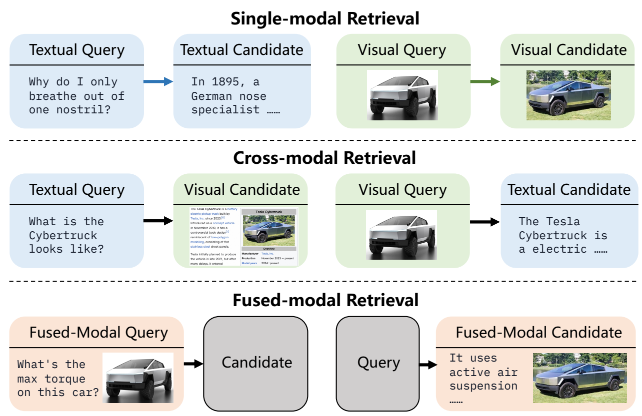

Figure 1. Illustration of different retrieval settings in our universal multimodal retrieval task. Blocks with black borders represent data in arbitrary modalities, *i.e*. text-only, image-only or fused.

Recently, researchers have turned to exploring Multimodal Large Language Models (MLLMs) [35, 65] in UMR. For example, it is shown that training MLLMs with text data alone can generate universal multimodal embeddings with respectable retrieval performance [22]. However, modality-limited training may fail to fully demonstrate the potential of MLLMs in UMR. We believe that incorporating multimodal data composition (as shown in Figure 1) could further enhance the model performance and generalization. Moreover, visual documents (*i.e*. document screenshots) are

*Equal Contribution. Work done during the internship of XZ, WX and ZD. DL is the tech lead. †Correspondence: mason.zms@gmail.com

Methods	Modeling		Retrieval Setting		
	Approach	Training	S&C	Fused	VD
UniVL-DR [39]	CLIP Feat. Fusion	Cross-modal	✓	✗	✗
UniIR [66]	CLIP Score Fusion BLIP Feat. Fusion	Multimodal	✓	✓	✗
MARVEL [75]	Text Enc.+Plugin	Cross-modal	✓	✗	✗
VISTA [74]	Text Enc.+Plugin	Multimodal	✓	✓	✗
E5-V [22]	MLLM	Text-only	✓	✓	✗
GME (Ours)	MLLM	Multimodal	✓	✓	✓

Table 1. Comparison of UMR studies. `Feat.` and `Enc.` are abbreviations for "Feature" and "Encoder". `S&C`, `Fused`, and `VD` denote the retrieval setting of single-modal & cross-modal, fused-modal, and retrieving visual documents (*e.g.* PDF screenshots), respectively. The setting explaination is in Figure 1.

increasingly important in UMR tasks, as they not only simplify the pipelines of diverse Retrieval-Augmented Generation (RAG) applications, but also mitigate information loss during modality conversion [12, 41]. However, current UMR models primarily target natural images, neglecting support for this scenario (Table 1).

To address the aforementioned challenges, we propose the General Multimodal Embedder (**GME**), an instruction-based embedding framework utilizing MLLMs as the backbone. GME enables retrieval across various modalities in the unified paradigm, including text, images, visual documents, and fused-modal[1] (*i.e.* image-text composed) contents. Our framework is underpinned by two key techniques: (1) A strategically optimized training data composition for UMR. We categorize UMR tasks into three types: single-modal, cross-modal, and fused-modal (Figure 1). Through extensive experimentation, we analyze how different compositions affect performance (Figure 3) and demonstrate that a balanced mixture of all types yields optimal results. (2) An efficient fused-modal data synthesis pipeline. Recognizing the under-representation of fused-modal data and its potential impact on training effectiveness, we develop a streamlined data synthesis pipeline (§4.2). This approach has successfully generated a comprehensive dataset of 1.1M fused-modal pairs, significantly enhancing our training and model capabilities.

To evaluate the effectiveness of our framework, we compile a comprehensive UMR Benchmark, namely **UMRB**. This benchmark encompasses tasks from widely recognized retrieval benchmarks in text [55], multimodal [66], and visual document retrieval [12], as well as our newly processed fused-modal retrieval data. We build our models on top of the strong Qwen2-VL series MLLMs [65] and train them on our constructed dataset. Experimental results demonstrate that our model achieves state-of-the-art performance on UMRB. Additionally, we perform in-depth analyses on model scaling, training strategies, and ablation of our synthetic data. Our key contributions are:

- We explore strategies to adapt MLLMs into UMR models, and present `GME`, a powerful embedding model capable of retrieving candidates across different modalities. `GME` is the first UMR model to deliver visual document retrieval performance on par with specialized models.
- We propose a novel data synthesis pipeline for constructing large-scale, fused-modal training data to encounter the scarcity of such training data. This pipeline is more efficient than previous approaches and can be easily extended to other domains.
- We compile the UMR benchmark, `UMRB`, to evaluate a broader range of retrieval tasks compared to existing benchmarks. `UMRB` categorizes tasks into three types: single-modal, cross-modal, and fused-modal, and offers a comprehensive performance evaluation across them.

2. Related Work

Multimodal Large Language Models The emergence of Large Language Models (LLMs) has driven significant progress in natural language processing [3, 49], leading to the development of Multimodal LLMs that extend these capabilities to handle multimodal information. Prominent MLLMs such as GPT-4V [48], LLaVa [35, 36], Qwen-VL [65], InternVL [7] and MiniCPM-V [71] have shown promising advancements in multimodal information understanding and reasoning. Typically, an MLLM consists of an LLM, a vision encoder, and a projector that bridges the two components by transforming raw multimodal inputs into vectors compatible with the LLM [72].

Multimodal Retrieval Early multimodal retrieval tasks focused on single-modal [73] or cross-modal retrieval [61]. Recently, the expansion of multimedia applications and multimodal retrieval-augmented generation (RAG) by MLLMs has created a need for unified multimodal retrieval models for complex scenarios. Existing approaches largely utilize pre-trained models such as CLIP [51] or BLIP [29] for multimodal embedding. For instance, UniVL-DR [39] and UniIR [66] initially encode images and texts separately using CLIP or BLIP encoders, followed by fusion strategies like score fusion to integrate features from both modalities. Additionally, VISTA [74] and MARVEL [75] employ pre-trained text embedding models enhanced with visual plugins to encode composite image-text candidates. However, these methods are typically designed for specific tasks like multimodal document retrieval and lack flexibility to handle diverse multimodal retrieval tasks.

Concurrent with our work, E5-V [22] and VLM2VEC [23] propose fine-tuning MLLMs on single-text (NLI [14]) or vision-centric relevance data, demonstrating

[1]We use fuse-modal instead of multimodal to denote the data that contains both text and image to disambiguate from the UMR task.

their transferability to multimodal retrieval. In this paper, we are the first to explore the fine-tuning of an MLLM-based universal multimodal retriever that can address both visual retrieval tasks and maintain strong text-to-text retrieval capabilities. Moreover, we are the first to extend a unified retrieval model to handle not only natural image retrieval but also text-rich image retrieval [12].

Embedding Models with Pre-trained Language Models With the advancement of pre-trained Language Models, research in both pure text and Vision-Language Models has focused on building representation models based on these pre-trained language models. In the text retrieval domain, state-of-the-art text embedding models such as Contriver [21], E5 [62], GTE [31], and BGE [68] are all built upon pre-trained language models and have demonstrated impressive generalization and robust performance in text retrieval tasks. Recently, leveraging LLMs combined with supervised fine-tuning (SFT), researchers have developed unified text representation models that fully utilize the text understanding capabilities of LLMs, resulting in models with enhanced performance and generalization [28, 31, 63]. These models typically process user text inputs through LLMs, using the hidden states from the final transformer layer—either through pooling or by selecting the last token—as the final representation. Inspired by the success of universal text embedding models based on text LLMs, researchers have begun to explore the construction of unified multimodal retrieval models using MLLMs [22, 23]. In this paper, we aim to demonstrate through systematic experiments that constructing a truly universal multimodal retrieval model using MLLMs is feasible.

3. Universal Multimodal Retrieval

Current UMR sub-tasks can be categorized into three types based on the modalities of the query and the candidate:

- **Single-Modal Retrieval**: Both the query and the candidate belong to the same modality, such as text-to-text (T→T) or image-to-image (I→I) retrieval scenarios.
- **Cross-Modal Retrieval**: The query and the candidate belong to different modalities, typically text-to-image (T→I) retrieval. Unlike most prior work that focuses on natural-style image retrieval, we also consider the retrieval of rich-text images (e.g., images converted from scholarly PDFs). We denote this scenario as text-to-visual document (T→VD) retrieval.
- **Fused-Modal Retrieval**: More complicated retrieval tasks involve mixed modalities in queries, candidates, or both. For example, in EVQA [46], both queries and candidates combine text and images.

The visualization of these settings refers to Figure 1.

[2]More details can be found at Quora Dataset Release: Question Pairs.

Class	Task	Datasets
Single-Modal (17)	T→T (16)	ArguAna[59] Climate-FEVER[11] CQADupStack[18] DBPedia[17] FEVER[56] FiQA2018[42] HotpotQA[70] MSMARCO[47] NFCorpus[2] NQ[26] Quora[2] SCIDOCS[8] SciFact[60] Touche2020[1] TRECCOVID[58] WebQA[4]
	I→I (1)	Nights[13]
Cross-Modal (18)	T→I (4)	VisualNews[34] Fashion200k[16] MSCOCO[32] Flickr30k[50]
	T→VD (10)	TAT-DQA[76] ArxivQA[30] DocVQA[44] InfoVQA[45] Shift Project† Artificial Intelligence† Government Reports† Healthcare Industry† Energy † TabFQuad†
	I→T (4)	VisualNews[34] Fashion200K[16] MSCOCO[32] Flickr30k[50]
Fused-Modal (12)	T→IT (2)	WebQA[4] EDIS[37]
	IT→T (5)	OVEN[20] INFOSEEK[6] ReMuQ[40] OKVQA[43] LLaVA[33]
	IT→I (2)	FashionIQ[67] CIRR[38]
	IT→IT (3)	OVEN[20] EVQA[46] INFOSEEK[6]

Table 2. An overview of tasks and datasets in our UMRB. † means that they all originate from [12].

3.1. Universal Multimodal Retrieval Benchmark

Based on the aforementioned classification principles, we introduce a new benchmark to comprehensively assess the performance of UMR models. This benchmark comprises **47** evaluation datasets that cover a broad spectrum of multimodal retrieval tasks, and we name it the Universal Multimodal Retrieval Benchmark (UMRB). These evaluation datasets primarily originate from previously constructed datasets tailored for each sub-scenario or sub-task. Specifically, UMRB includes: (1) The BEIR [55] benchmark for text-to-text retrieval scenarios; (2) The M-BEIR [66] dataset for vision-centric retrieval scenarios; (3) Additional fused-modal datasets that not cover by M-BEIR; and (4) text-to-visual document search datasets, such as ViDoRe [12], to extend the coverage of our benchmark and ensure a comprehensive evaluation of model universality. A detailed list of the UMRB datasets is presented in Table 2.

Given the extensive size of UMRB, to expedite our experimental validation and analysis, we have sampled a subset of datasets from each category, constituting a smaller dataset named UMRB-Partial. This subset retains 39% of the total datasets while maintaining evaluation richness. More detailed statistical information about UMRB-Partial can be found in our supplementary materials.

4. Method

In this section, we present the training framework for developing the General Multimodal Embedder (GME) model. We describe the contrastive learning approach used to train the embedding model. Building on this, we conduct detailed experiments to determine the optimal balance of

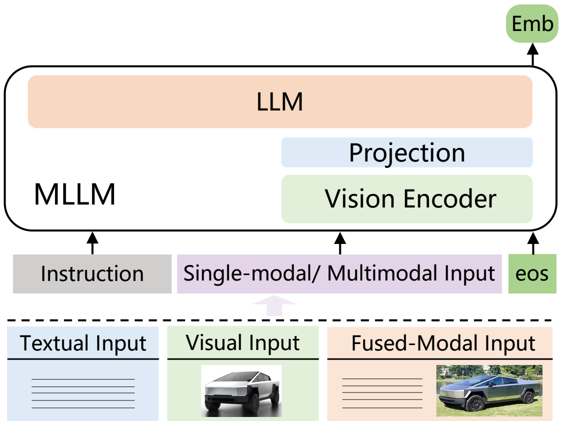

Figure 2. The GME model architecture. Emb denotes the embedding of the input content.

training data type. Specifically, our experiments demonstrate that diverse data type mixtures significantly enhances the model's ability to perform retrieval across various modalities. Lastly, recognizing the scarcity of high-quality fused-modal training data, we propose a novel method for automatically synthesizing large-scale, high-quality training data using MLLM.

4.1. GME: General Multimodal Embedder

Model Architecture We employ a MLLM as the foundation for GME. This model can accept images, text, or image-text pairs as input. Inspired by previous research on text embedding [31, 63], we use the final hidden state of the last token as the representation (or embedding) for the input. Although pre-trained MLLMs possess strong multimodal understanding capabilities, their original training objectives are not optimized for representation learning. Therefore, task-specific fine-tuning (or alignment) is necessary to enhance the model's representational capacity. Contrastive learning has been shown to effectively train LLMs and MLLMs to produce retrieval embeddings [22, 31].

Contrastive Learning In our contrastive learning setup, each training instance comprises a query q, a relevant candidate c, and a set of irrelevant candidates $\{c_1^-, c_2^-, \ldots, c_K^-\}$. Both q and c can be text, images, or image-text pairs, allowing the model to handle diverse data modalities. To tailor the model to various downstream retrieval tasks, we incorporate an instruction tuning method by including a tailored instructional text i with each retrieval task. For example, for the Visual Question Answering (VQA) task, the instruction could be: "Retrieve a passage that provides an answer to the given query about the image" guiding the model on how to process and interpret the query for specific objectives.

During training, we input q and instruction i into the model to obtain the query representation e_q. Similarly, each candidate c is input into the model to obtain its representation e_c. The training objective minimizes the cosine distance between e_q and e_c for relevant pairs while maximizing the distance between e_q and e_{c^-} for irrelevant pairs. Cosine similarity is employed to measure the directional alignment between embeddings, effectively capturing semantic similarities irrespective of their magnitudes.

The optimization process utilizes the InfoNCE loss function [57], defined as:

$$\mathcal{L} = -\log \frac{\exp\left(cos(e_q, e_c^+)/\tau\right)}{\exp\left(cos(e_q, e_c^+)/\tau\right) + \sum_{i=1}^{K} \exp\left(cos(e_q, e_{c_i^-})/\tau\right)}$$

where τ is the temperature parameter that scales the cosine similarities to control the distribution's concentration. This approach ensures that the model effectively learns to distinguish relevant from irrelevant information across different modalities, thereby enhancing its performance in multimodal retrieval tasks.

Hard Negatives The quality and diversity of negative samples are essential for improving contrastive learning [53]. Inspired by ANCE [69], we employ a two-stage training strategy: (1) Initial Training: We first train the model using randomly selected negative candidates, resulting in Model M_1. (2) Hard Negative Mining and Continue Training: Using M_1, we retrieve the top K candidates for each query and select non-relevant candidates from them as hard negatives. We then use these hard negatives to further train M_1, refining it into the final model. This ensures that the model can learn from both easily distinguishable and more challenging examples, thereby enhancing performance.

Training Data Composition A critical factor in multimodal representation learning is the composition of training data. Although previous studies like [22] have demonstrated that MLLMs can develop multimodal representation capabilities after being fine-tuned on single-modal data, the effect of data diversity on model performance remains unclear. Therefore, we compare the performance of models trained with different data combinations across various retrieval scenarios within our classification principle. Specifically, we used four types of training data: single-modal (including T→T and I→I), cross-modal (including T→VD and T→I), fused-modal training data (including IT→IT), and a mixed dataset combining the first three types. These different training data types result in a total of six models.

For single-modal data, we utilized the T→T dataset from MSMARCO [47] and the I→I dataset from ImageNet [10], treating images within the same category as positive matches and those from different categories as negatives. For cross-modal data, we employed T→I pairs

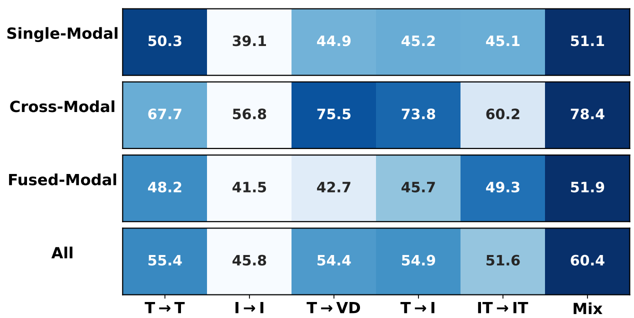

Figure 3. Impact of training data on multimodal retrieval tasks.

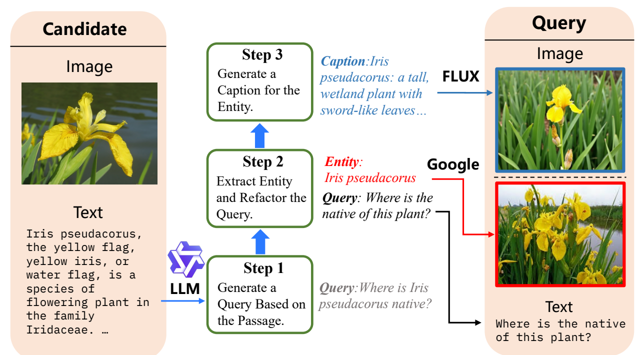

Figure 4. Pipeline for synthesizing fused-modal training data.

from the LAION [54] dataset and T→VD pairs from the Docmatix [27] dataset. For fused-modal data, we use the EVQA [46] dataset (IT→IT). For each subcategory, we randomly sampled 100,000 training instances to train the models independently. For the mixed dataset, we uniformly sampled 20,000 instances from each of the five datasets to train the final model, ensuring fair and reliable comparative experimental results. The performance of these six models on the UMRB-Partial test dataset is presented in Figure 3.

The results indicate that: (1) Models trained on single data types excel in corresponding retrieval tasks. For instance, models trained on T→T data performed best in text retrieval tasks.[3] (2) A balanced mix of different data types enhanced performance across various settings. This suggests that increasing the diversity of training modalities effectively improves the model's overall retrieval capabilities.

The above analysis highlights the importance of adequately representing each data type in training datasets to develop models that meet the requirements of universal multi-modal retrieval. During data collection, we observed that single-modal and cross-modal data are abundant, with over ten million training instances available. In contrast, fused-modal data remains limited. Common fused-modal training datasets such as EVQA[46], INFOSEEK[6], and CIRR [38] collectively contain fewer than one million instances. Additionally, these existing fused-modal datasets cover only a limited range of domains. Thus, efficiently supplementing high-quality fused-modal training data is essential. To address this challenge, we propose leveraging the generative capabilities of LLMs and MLLMs to synthesize additional training data.

4.2. Fused-Modal Data Synthesis

To efficiently synthesize high-quality data while minimizing manual intervention, we adopt a strategy similar to Doc2Query [15]. However, our approach differs in that we aim to generate fuse-modal candidate-to-query relevance data instead of single-modality, text-based relevance pairs. This requires obtaining high-quality candidates that include both image and text content. We primarily extracted such data from Wikipedia paragraphs[4]. Additionally, to enhance the domain diversity of the candidate data, we employed a domain classification model[5] to perform fine-grained classification of Wikipedia data into categories such as animals and plants. We then uniformly sampled from these categories and retained data with classification confidence scores above 0.5. Ultimately, we obtained 313,284 candidate entries, each containing both text and image content.

Based on the prepared data, the overall synthesis pipeline (Figure 4) could be divided into the following steps:

• **Doc2Query Generation**: The passage content from each candidate is input into an LLM[6] using a prompt to generate a natural query. To ensure the quality of the generated queries, we built a vector index of all passage contents using a text vector retrieval model[7]. Each generated query is then used to retrieve the corresponding passage from this collection. If the passage associated with the query is not within the top 20 retrieved items, the query is considered low quality due to low relevance and is discarded. In this step, we discarded 1.2% of the total generated queries. This process allows us to construct T→IT training data.

• **Entity Extraction and Query Rewrite**: We aim for the synthesized queries to include both texts and images (i.e., IT→IT type). To achieve this, we leverage entity extraction followed by image retrieval for the extracted entities and caption generation to supplement the image data on the query side. Specifically, for each generated query q from the first step, we prompt the LLM to extract entities from it with the text passage as reference, and then rewrite the original query into q'. For example, the query "Where is Iris pseudacorus native?" is transformed by the model to the rewritten query "Where is the native habitat of this plant?" with the entity "Iris pseudacorus" extracted. We then seek images

[3]Detail results are shown in the supplementary materials.

[4]github.com/google-research-datasets/wit/blob/main/wikiweb2m.md

[5]hf.co/facebook/bart-large-mnlifacebook

[6]In the entire pipeline, we utilize Qwen2.5-72B-Instruct as our LLM.

[7]hf.co/Alibaba-NLP/gte-Qwen2-1.5B-instruct

that match this entity and combine them with the rewritten query q' to form the final fuse-modal query.

• **Image Retrieval and Generation**: We explore two methods for obtaining images. The first method uses the Google Image Search API[8] to retrieve images matching the entity terms, retaining the top five results. The second method involves generating images using a text-to-image model[9]. Specifically, we first use the LLM to generate a caption suitable for image generation based on the entity and the passage of the generated query, then input this caption into the text-to-image generation model to create the corresponding image. This approach allows us to quickly and efficiently obtain high-quality, diverse images. The synthesized results can also be assembled into IT→ IT retrieval type data.

• **Data Filtering**: To ensure the quality of the synthesized data, we perform filtering [9] on the final dataset. We observe that images generated by the FLUX model have consistent quality, whereas images retrieved via the Google Image Search API often include noisy data. Therefore, for images obtained through the Google Image Search API, we use the CLIP model[10] to assess image-caption relevance. Images with a relevance score below 0.2 were filtered out.

Through the synthesis pipeline, we produce 1,135,000 high-quality fuse-modal training data entries (including T→IT and IT→IT types). After filtering, we retain 1,102,000 entries, resulting in a data loss rate of 2.9%. The entire process consumed 600 A100 GPU hours. Detailed descriptions of all prompts used in the data synthesis pipeline and examples of the synthesized data are provided in the supplementary material.

5. Experiments

5.1. Settings

Training Data Building on the findings from §4.1, we train our model using a diverse dataset of **8** million instances spanning various retrieval modalities. For single-modal retrieval tasks, we utilize datasets including MS-MARCO [47], NQ [26], HotpotQA [70], TriviaQA [24], SQuAD [52], FEVER [56], and AllNLI for SimCSE [14], selecting a total of 1 million entries. From ImageNet [10], we extract 1 million image-to-image training instances, designating images within the same class as positive samples and others as negative samples. For cross-modal retrieval tasks, we incorporate 2 million entries from the LAION [54], MSCOCO [32], and Docmatix [27] datasets. Additionally, for fused-modal retrieval tasks, we include a total of 2 million instances: 1.1 million synthesized by us, and the remaining from the M-BEIR [66] training data.

[8] https://serpapi.com/google-images-api
[9] https://hf.co/black-forest-labs/FLUX.1-dev
[10] https://hf.co/openai/clip-vit-large-patch14

Training Configuration We use Qwen2-VL [65] model series as the backbone for our MLLM, conducting training on models with both 2 billion (2B) and 7 billion (7B) parameters. Our training utilizes Low-Rank Adaptation (LoRA) [19] with a rank of 8, a learning rate of 1e-4, and a temperature setting of 0.03. To manage the varying number of visual tokens required by Qwen2-VL for different image resolutions and maintain training efficiency, we limit the maximum number of visual tokens per image to 1,024.

For data with images, we set the maximum text length to 1,800 tokens, using a batch size of 128 for the 2B model and 32 for the 7B model. For text-only data, the maximum length was set to 512 tokens, with batch size of 512 for the 2B model and 128 for the 7B model. Each training sample included 8 negative examples. To conserve GPU memory, we employ gradient checkpointing [5] and train the model using bfloat16 precision. All training was conducted on eight NVIDIA A100 GPUs, each with 80GB of memory.

Baselines We compare our method against four types of retrieval systems: (1) Previous representative UMR models, for example, VISTA [74] for text encoder based, and E5-V [22] for MLLM based; (2) Powerful multimodal representation (embedding) models, *i.e.* One-Peace [64], which supports modalities beyond text and image and hence could also be tested on our `UMRB`; (3) Recent visual document retrieval models, namely DSE [41]; and (4) the classic cross-modal pipeline, CLIP score-fusion, denoted as CLIP-SF, which provides top-tier cross-modal performance. We exclude comparisons with state-of-the-art text retrieval models as VISTA demonstrates comparable performance levels.

5.2. Main Results

Table 3 presents the evaluation results of the baseline systems alongside our proposed `GME`. Scores are averaged across each sub-task and categorized by retrieval modality type: single-modal, cross-modal, and fused-modal. Additionally, the overall micro-average score on the `UMRB` is in the last column. First, focusing on the average scores, our smaller model, *i.e.* `GME-Qwen2-VL-2B`, already outperforms the previous state-of-the-art UMR model (VISTA [74]). The larger model, *i.e.* `GME-Qwen2-VL-7B`, further enhances this performance, demonstrating the effectiveness of our approach in handling UMR tasks.

Second, our models outperform smaller methods such as VISTA (million-level parameters) and One-Peace (4B parameters). The larger MLLM baseline, E5-V [22] (8B parameters), performs well in text-dominated tasks (e.g., T→T) but falls short in other areas. This indicates that training with multimodal data is crucial for achieving superior performance in UMR tasks. Our training data provides a stronger foundation for future advancements.

Next, the cross-modal pipeline CLIP-SF outperforms

UMRB	Size	Single-Modal		Cross-Modal			Fused-Modal				Avg.
Task (#Datasets)		T→T (16)	I→I (1)	T→I (4)	T→VD (10)	I→T (4)	T→IT (2)	IT→T (5)	IT→I (2)	IT→IT (3)	(47)
VISTA [74]	0.2B	55.15	**31.98**	32.88	10.12	31.23	45.81	53.32	8.97	26.26	37.32
CLIP-SF [66]	0.4B	39.75	31.42	59.05	24.09	62.95	66.41	53.32	34.90	55.65	43.66
One-Peace [64]	4B	43.54	31.27	61.38	42.9	65.59	42.72	28.29	6.73	23.41	42.01
DSE [41]	4.2B	48.94	27.92	40.75	78.21	52.54	49.62	35.44	8.36	40.18	50.04
E5-V [22]	8.4B	52.41	27.36	46.56	41.22	47.95	54.13	32.90	23.17	7.23	42.52
GME-Qwen2VL-2B	2.2B	55.93	29.86	57.36	87.84	61.93	76.47	64.58	37.02	66.47	64.45
GME-Qwen2VL-7B	8.2B	**58.19**	31.89	**61.35**	**89.92**	**65.83**	**80.94**	**66.18**	**42.56**	**73.62**	**67.44**

Table 3. Results of different models on our benchmark. Following previous works [12, 55, 66], we present NDCG@10 scores for T→T tasks, excluding the WebQA dataset. For T→VD tasks, we provide NDCG@5 scores. For the Fashion200K, FashionIQ and OKVQA datasets, we report Recall@10 scores, while for all other datasets, we report Recall@5 scores.

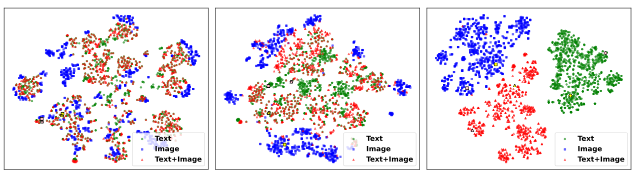

Figure 5. Visualization of the embeddings in a 2D plot by T-SNE. Left: Our GME, Middle: VISTA, Right: CLIP. We use instances from Encyclopedia VQA and highlight two semantic groups with yellow and pink labels, respectively. Please zoom in to view them.

Setting	Single	Cross	Fused	Average
w/ EVQA	45.13	60.21	49.32	51.55
w/ Gen_{Flux}	46.27	61.19	51.46	52.97
w/ $\text{Gen}_{\text{Google}}$	47.08	61.35	52.01	53.48

Table 4. Results of GME-Qwen2-VL-2B trained with different generated datasets and evaluated on UMRB-Partial.

UMR models like VISTA, E5-V, and One-Peace. For VISTA and E5-V, the performance gap is likely due to limitations in their text-modality bounds: VISTA is constrained by the text embedding space of its fixed backbone, and E5-V is limited by text-only training. One-Peace's modality alignment-centered modeling may not be optimized for fused-modal content. In contrast, our models are specifically designed to handle fused-modal data, resulting in significantly better performance compared to the baselines. Although our training data includes several previously constructed fused-modal datasets, the contribution of our generated fused-modal training data will be discussed in §5.3.

Finally, we compare with the recent visual document retrieval model DSE [41], specialized for the T→VD task within the Cross-Modal group, which has approximately 4B parameters. Our models are competitive with or exceed the performance of this task-specific baseline, demonstrating the feasibility and promise of integrating visual document retrieval into a unified retriever framework.

5.3. Analyses

Are the Produced Embeddings Modality Universal? Given our the impressive performance of our model, we assess the quality of its embeddings. Specifically, we investigate whether the embeddings are modality-universal meaning that embeddings representing the same semantic content across different modalities are closely clustered in the embedding space, or if they remain in separate sub-spaces tailored for each modality-specific task. To probe this question, we sample 1000 instances from the EVQA dataset and visualize their embeddings of different modalities by t-SNE, as shown in Figure 5. We also highlight two semantic close groups with yellow and pink labels, respectively. We can observe that the embeddings from CLIP are distinctly separated by modality, whereas the embeddings from our model are intermingled and organized semantically. Meanwhile, the points from the same semantic group are closely clustered. This demonstrates that our model effectively generates modality-universal representations, enhancing its applicability across various UMR tasks.

Ablation Study on Synthetic Fused-Modal Data We propose an efficient data synthesis pipeline (§4.2) and generate large-scale fused-modal pairs to support model training. After witnessing the state-of-the-art performance of our model, it is natural to question the contribution of this synthetic data to the overall performance. To this end, we conduct an ablation study using three parallel training datasets, each comprising 100,000 pairs: original EVQA data, synthetic data with Google-retrieved images ($\text{Gen}_{\text{Google}}$), and synthetic data with FLUX-generated images (Gen_{Flux}). We train three models with identical parameters on these datasets and evaluate their performance on UMRB-Partial, with results shown in Table 4. Both synthetic datasets outperform the original EVQA data, indicating the high quality of our synthesized data. Although

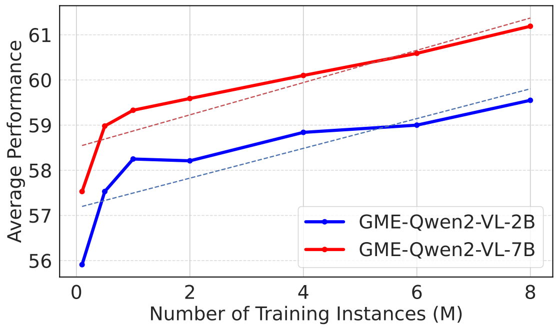

Figure 6. Average Performance of GME-Qwen2-VL-2B (Blue) and GME-Qwen2-VL-7B (Red) on UMRB-Partial, trained with varying numbers of training instances.

Google-retrieved images achieved marginally better performance than FLUX-generated images, the difference is minor and acceptable given the potential limitations of the Google Search API for rapid, large-scale dataset generation.

Training Scaling Law Our approach is primarily data-centric, constructing a diverse training dataset of approximately 8 million samples across various UMR settings (§5.1). Training on such a large-scale dataset demands significant computational resources and time. Therefore, we explored the training scaling law by examining how model performance evolves with increasing training steps. Due to the time-consuming nature of evaluating certain retrieval tasks, we assessed performance on our `UMRB-Partial` dataset for faster evaluation. Figure 6 illustrates the performance progression of our 2B and 7B models on `UMRB-Partial` during training. Both models exhibit linear performance improvements as training continues, suggesting that extended training could yield further benefits. However, due to time constraints, we halted current training. Future work will investigate longer training periods to enhance model performance further.

Ablation Study on Modeling We conduct an ablation study to investigate the effectiveness of different design choices of `GME`. We consider the following three aspects: (1) Fine-tuning strategy. Our final models are trained by LoRA with rank 8. We compare with other rank values and full fine-tuning. The results in the first group of Table 5 show that LoRA with rank 8 yields the best performance. (2) Training data organization. We compare models trained without hard negative mining. The second group of Table 5 demonstrates that the removal of hard negatives led to performance declines, indicating that it is essential for effective retrieval model training. (3) Retrieval instructions. We compare models trained without retrieval instructions. The third group shows that retrieval instructions are crucial for better UMR. (4) Modeling techniques. Our final models are in the casual attention mode and use the EOS token state as the embedding, hence we compare the performance of the model trained with mean pooling and the bi-directional attention mechanism. The last group of Table 5 shows that these alternative settings negatively impact performance.

Setting	Single	Cross	Fused	Average
Fine-tuning strategy				
LoRA r=8	**48.09**	78.39	**51.88**	**59.45**
LoRA r=16	47.86	**78.63**	51.42	59.30
LoRA r=32	47.85	78.55	50.48	58.96
LoRA r=64	47.65	78.61	51.09	59.11
Full training	43.16	75.79	49.28	56.07
Training data organization				
w/o hard-negative	47.55	78.01	50.95	58.83
Retrieval Setting				
w/o Instruction	46.82	78.10	49.09	58.00
Model Design				
w/ mean pooling	47.86	77.95	51.33	59.04
w/ bi-attention	46.55	76.78	49.54	57.62

Table 5. Results of the ablation study on Qwen2-VL-2B. All models are trained using 100,000 instances, consistent with the experimental setup described in Section 4.1.

6. Conclusion

In this work, we target the universal multimodal retrieval (UMR) problem. We begin by systematically categorizing current UMR tasks, proposing a comprehensive classification framework. Based on this, we explore ways to further improve MLLM-based UMR models, suggesting the GME model. The GME models are trained using contrastive learning loss on a diverse set of multimodal data settings, while also extending support for visual retrieval. Additionally, to overcome limitations in existing UMR evaluation benchmarks, we compiled a new comprehensive benchmark (i.e., UMRB) by integrating multiple data sources. This benchmark effectively balances existing UMR tasks with the increasingly important text and visual document retrieval tasks, enabling a more thorough assessment of UMR model performance. We evaluate existing UMR models and our proposed GME model on UMRB, finding that our model achieves state-of-the-art performance. We also conducted various analyses to validate the effectiveness of our methods and enhance our understanding of them. Our benchmark, models, and other materials are open-source at `https://hf.co/Alibaba-NLP/gme-Qwen2-VL-7B-Instruct`.

Acknowledgments

This work receives partial support from the Natural Science Foundation of China (under Grant 624B2048) and Research Grant Council of Hong Kong (PolyU/15209724).

References

[1] Alexander Bondarenko, Maik Fröbe, Meriem Beloucif, Lukas Gienapp, Yamen Ajjour, Alexander Panchenko, Chris Biemann, Benno Stein, Henning Wachsmuth, Martin Potthast, and Matthias Hagen. Overview of touché 2020: Argument retrieval - extended abstract. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction - CLEF 2020*, pages 384–395. Springer, 2020. 3

[2] Vera Boteva, Demian Gholipour Ghalandari, Artem Sokolov, and Stefan Riezler. A full-text learning to rank dataset for medical information retrieval. In *Advances in Information Retrieval - 38th European Conference on IR Research, ECIR 2016, Padua, Italy*, pages 716–722. Springer, 2016. 3

[3] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020. 2

[4] Yingshan Chang, Guihong Cao, Mridu Narang, Jianfeng Gao, Hisami Suzuki, and Yonatan Bisk. Webqa: Multihop and multimodal QA. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16474–16483, 2022. 1, 3

[5] Tianqi Chen, Bing Xu, Chiyuan Zhang, and Carlos Guestrin. Training deep nets with sublinear memory cost. *CoRR*, abs/1604.06174, 2016. 6

[6] 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 *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 14948–14968, Singapore, 2023. Association for Computational Linguistics. 3, 5

[7] Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *Science China Information Sciences*, 67(12):220101, 2024. 2

[8] Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel Weld. SPECTER: Document-level representation learning using citation-informed transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2270–2282, Online, 2020. Association for Computational Linguistics. 3

[9] Zhuyun Dai, Vincent Y. Zhao, Ji Ma, Yi Luan, Jianmo Ni, Jing Lu, Anton Bakalov, Kelvin Guu, Keith B. Hall, and Ming-Wei Chang. Promptagator: Few-shot dense retrieval from 8 examples. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. 6

[10] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, Miami, USA*, pages 248–255. IEEE Computer Society, 2009. 4, 6

[11] Thomas Diggelmann, Jordan L. Boyd-Graber, Jannis Bulian, Massimiliano Ciaramita, and Markus Leippold. CLIMATE-FEVER: A dataset for verification of real-world climate claims. *CoRR*, abs/2012.00614, 2020. 3

[12] Manuel Faysse, Hugues Sibille, Tony Wu, Bilel Omrani, Gautier Viaud, Céline Hudelot, and Pierre Colombo. Colpali: Efficient document retrieval with vision language models. In *The Thirteenth International Conference on Learning Representations*, 2025. 2, 3, 7

[13] Stephanie Fu, Netanel Yakir Tamir, Shobhita Sundaram, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. Dreamsim: Learning new dimensions of human visual similarity using synthetic data. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 3

[14] Tianyu Gao, Xingcheng Yao, and Danqi Chen. SimCSE: Simple contrastive learning of sentence embeddings. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic, 2021. Association for Computational Linguistics. 2, 6

[15] Mitko Gospodinov, Sean MacAvaney, and Craig Macdonald. Doc2query-: When less is more. In *Advances in Information Retrieval - 45th European Conference on Information Retrieval, ECIR 2023*, pages 414–422, Dublin, Ireland, 2023. Springer. 5

[16] Xintong Han, Zuxuan Wu, Phoenix X. Huang, Xiao Zhang, Menglong Zhu, Yuan Li, Yang Zhao, and Larry S. Davis. Automatic spatially-aware fashion concept discovery. In *IEEE International Conference on Computer Vision, ICCV 2017*, pages 1472–1480, Venice, Italy, 2017. IEEE Computer Society. 3

[17] 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 *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, page 1265–1268, New York, NY, USA, 2017. Association for Computing Machinery. 3

[18] Doris Hoogeveen, Karin M. Verspoor, and Timothy Baldwin. Cqadupstack: A benchmark data set for community question-answering research. In *Proceedings of the 20th Australasian Document Computing Symposium*, New York, NY, USA, 2015. Association for Computing Machinery. 3

[19] Edward J Hu, yelong shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen.

LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. 6

[20] 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 *IEEE/CVF International Conference on Computer Vision, ICCV 2023*, pages 12031–12041, Paris, France, 2023. IEEE. 3

[21] Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. Unsupervised dense information retrieval with contrastive learning. *Transactions on Machine Learning Research*, 2022. 3

[22] 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. *CoRR*, abs/2407.12580, 2024. 1, 2, 3, 4, 6, 7

[23] Ziyan Jiang, Rui Meng, Xinyi Yang, Semih Yavuz, Yingbo Zhou, and Wenhu Chen. VLM2vec: Training vision-language models for massive multimodal embedding tasks. In *The Thirteenth International Conference on Learning Representations*, 2025. 2, 3

[24] Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada, 2017. Association for Computational Linguistics. 6

[25] 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 *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online, 2020. Association for Computational Linguistics. 1

[26] 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*, 7:452–466, 2019. 3, 6

[27] Hugo Laurençon, Andrés Marafioti, Victor Sanh, and Leo Tronchon. Building and better understanding vision-language models: insights and future directions. In *Workshop on Responsibly Building the Next Generation of Multimodal Foundational Models*, 2024. 5, 6

[28] 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. In *The Thirteenth International Conference on Learning Representations*, 2025. 3

[29] Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. BLIP: bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning, ICML 2022*, pages 12888–12900, Baltimore, Maryland, USA, 2022. PMLR. 2

[30] Lei Li, Yuqi Wang, Runxin Xu, Peiyi Wang, Xiachong Feng, Lingpeng Kong, and Qi Liu. Multimodal ArXiv: A dataset for improving scientific comprehension of large vision-language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14369–14387, Bangkok, Thailand, 2024. Association for Computational Linguistics. 3

[31] Zehan Li, Xin Zhang, Yanzhao Zhang, Dingkun Long, Pengjun Xie, and Meishan Zhang. Towards general text embeddings with multi-stage contrastive learning. *CoRR*, abs/2308.03281, 2023. 3, 4

[32] Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In *13th European Conference on Computer Vision, ECCV 2014*, pages 740–755, Zurich, Switzerland, 2014. Springer. 3, 6

[33] Weizhe Lin, Jingbiao Mei, Jinghong Chen, and Bill Byrne. PreFLMR: Scaling up fine-grained late-interaction multimodal retrievers. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5294–5316, Bangkok, Thailand, 2024. Association for Computational Linguistics. 3

[34] Fuxiao Liu, Yinghan Wang, Tianlu Wang, and Vicente Ordonez. Visual news: Benchmark and challenges in news image captioning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6761–6771, Online and Punta Cana, Dominican Republic, 2021. Association for Computational Linguistics. 3

[35] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. 1, 2

[36] Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. Llava-next: Improved reasoning, ocr, and world knowledge, 2024. 2

[37] Siqi Liu, Weixi Feng, Tsu-Jui Fu, Wenhu Chen, and William Wang. EDIS: Entity-driven image search over multimodal web content. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4877–4894, Singapore, 2023. Association for Computational Linguistics. 3

[38] Zheyuan Liu, Cristian Rodriguez Opazo, Damien Teney, and Stephen Gould. Image retrieval on real-life images with pre-trained vision-and-language models. In *2021 IEEE/CVF International Conference on Computer Vision, ICCV 2021*, pages 2105–2114, Montreal, Canada, 2021. IEEE. 3, 5

[39] 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 *The Eleventh International Conference on Learning Representations*, 2023. 1, 2

[40] Man Luo, Zhiyuan Fang, Tejas Gokhale, Yezhou Yang, and Chitta Baral. End-to-end knowledge retrieval with multi-

modal queries. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8573–8589, Toronto, Canada, 2023. Association for Computational Linguistics. 3

[41] Xueguang Ma, Sheng-Chieh Lin, Minghan Li, Wenhu Chen, and Jimmy Lin. Unifying multimodal retrieval via document screenshot embedding. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6492–6505, Miami, Florida, USA, 2024. Association for Computational Linguistics. 2, 6, 7

[42] 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 of the The Web Conference 2018 on The Web Conference 2018*, pages 1941–1942, Lyon, France, 2018. ACM. 3

[43] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. OK-VQA: A visual question answering benchmark requiring external knowledge. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2019*, pages 3195–3204, Long Beach, CA, USA, 2019. Computer Vision Foundation / IEEE. 3

[44] Minesh Mathew, Dimosthenis Karatzas, and C. V. Jawahar. Docvqa: A dataset for VQA on document images. In *IEEE Winter Conference on Applications of Computer Vision, WACV 2021*, pages 2199–2208, Waikoloa, HI, USA, 2021. IEEE. 3

[45] Minesh Mathew, Viraj Bagal, Rubèn Tito, Dimosthenis Karatzas, Ernest Valveny, and C. V. Jawahar. Infographicvqa. In *IEEE/CVF Winter Conference on Applications of Computer Vision, WACV 2022, Waikoloa, HI, USA, January 3-8, 2022*, pages 2582–2591. IEEE, 2022. 3

[46] Thomas Mensink, Jasper R. R. Uijlings, Lluís Castrejón, Arushi Goel, Felipe Cadar, Howard Zhou, Fei Sha, André Araújo, and Vittorio Ferrari. Encyclopedic VQA: visual questions about detailed properties of fine-grained categories. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023*, pages 3090–3101, Paris, France, 2023. IEEE. 3, 5

[47] Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. MS MARCO: A human generated machine reading comprehension dataset. In *Proceedings of the Workshop on Cognitive Computation: Integrating neural and symbolic approaches 2016*, Barcelona, Spain, 2016. CEUR-WS.org. 3, 4, 6

[48] OpenAI. Gpt-4v(ision) system card, 2023. 2

[49] OpenAI. GPT-4 technical report. *CoRR*, abs/2303.08774, 2023. 2

[50] Bryan A. Plummer, Liwei Wang, Chris M. Cervantes, Juan C. Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *2015 IEEE International Conference on Computer Vision, ICCV 2015*, pages 2641–2649, Santiago, Chile, 2015. IEEE Computer Society. 3

[51] 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 *Proceedings of the 38th International Conference on Machine Learning, ICML 2021*, pages 8748–8763. PMLR, 2021. 2

[52] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. SQuAD: 100,000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, Texas, 2016. Association for Computational Linguistics. 6

[53] Joshua David Robinson, Ching-Yao Chuang, Suvrit Sra, and Stefanie Jegelka. Contrastive learning with hard negative samples. In *International Conference on Learning Representations*, 2021. 4

[54] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade W Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa R Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5b: An open large-scale dataset for training next generation image-text models. In *Thirty-sixth Conference on Neural Information Processing Systems: Datasets and Benchmarks Track*, 2022. 5, 6

[55] Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. BEIR: A heterogeneous benchmark for zero-shot evaluation of information retrieval models. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*, 2021. 2, 3, 7

[56] James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana, 2018. Association for Computational Linguistics. 3, 6

[57] Aäron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *CoRR*, abs/1807.03748, 2018. 4

[58] Ellen Voorhees, Tasmeer Alam, Steven Bedrick, Dina Demner-Fushman, William R. Hersh, Kyle Lo, Kirk Roberts, Ian Soboroff, and Lucy Lu Wang. Trec-covid: constructing a pandemic information retrieval test collection. *SIGIR Forum*, 54(1), 2021. 3

[59] Henning Wachsmuth, Shahbaz Syed, and Benno Stein. Retrieval of the best counterargument without prior topic knowledge. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 241–251, Melbourne, Australia, 2018. Association for Computational Linguistics. 3

[60] David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7534–7550, Online, 2020. Association for Computational Linguistics. 3

[61] Kaiye Wang, Qiyue Yin, Wei Wang, Shu Wu, and Liang Wang. A comprehensive survey on cross-modal retrieval. *CoRR*, abs/1607.06215, 2016. 2

[62] Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. Text embeddings by weakly-supervised contrastive pre-training. *CoRR*, abs/2212.03533, 2022. 3

[63] Liang Wang, Nan Yang, Xiaolong Huang, Linjun Yang, Rangan Majumder, and Furu Wei. Improving text embeddings with large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11897–11916, Bangkok, Thailand, 2024. Association for Computational Linguistics. 3, 4

[64] Peng Wang, Shijie Wang, Junyang Lin, Shuai Bai, Xiaohuan Zhou, Jingren Zhou, Xinggang Wang, and Chang Zhou. ONE-PEACE: exploring one general representation model toward unlimited modalities. *CoRR*, abs/2305.11172, 2023. 6, 7

[65] Peng Wang, Shuai Bai, Sinan Tan, Shijie Wang, Zhihao Fan, Jinze Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Yang Fan, Kai Dang, Mengfei Du, Xuancheng Ren, Rui Men, Dayiheng Liu, Chang Zhou, Jingren Zhou, and Junyang Lin. Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. *CoRR*, abs/2409.12191, 2024. 1, 2, 6

[66] 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. In *18th European Conference on Computer Vision*, page 387–404, Milan, Italy, 2024. Springer-Verlag. 1, 2, 3, 6, 7

[67] Hui Wu, Yupeng Gao, Xiaoxiao Guo, Ziad Al-Halah, Steven Rennie, Kristen Grauman, and Rogério Feris. Fashion IQ: A new dataset towards retrieving images by natural language feedback. In *IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2021*, pages 11307–11317. Computer Vision Foundation / IEEE, 2021. 3

[68] Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. C-pack: Packed resources for general chinese embeddings. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, page 641–649, New York, NY, USA, 2024. Association for Computing Machinery. 3

[69] 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 *9th International Conference on Learning Representations*, 2021. 4

[70] 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 *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium, 2018. Association for Computational Linguistics. 3, 6

[71] Yuan Yao, Tianyu Yu, Ao Zhang, Chongyi Wang, Junbo Cui, Hongji Zhu, Tianchi Cai, Haoyu Li, Weilin Zhao, Zhihui He, Qianyu Chen, Huarong Zhou, Zhensheng Zou, Haoye Zhang, Shengding Hu, Zhi Zheng, Jie Zhou, Jie Cai, Xu Han, Guoyang Zeng, Dahai Li, Zhiyuan Liu, and Maosong Sun. Minicpm-v: A GPT-4V level MLLM on your phone. *CoRR*, abs/2408.01800, 2024. 2

[72] Shukang Yin, Chaoyou Fu, Sirui Zhao, Ke Li, Xing Sun, Tong Xu, and Enhong Chen. A survey on multimodal large language models. *National Science Review*, 11(12), 2024. 2

[73] Wayne Xin Zhao, Jing Liu, Ruiyang Ren, and Ji-Rong Wen. Dense text retrieval based on pretrained language models: A survey. *ACM Trans. Inf. Syst.*, 42(4):89:1–89:60, 2024. 2

[74] Junjie Zhou, Zheng Liu, Shitao Xiao, Bo Zhao, and Yongping Xiong. VISTA: Visualized text embedding for universal multi-modal retrieval. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3185–3200, Bangkok, Thailand, 2024. Association for Computational Linguistics. 1, 2, 6, 7

[75] Tianshuo Zhou, Sen Mei, Xinze Li, Zhenghao Liu, Chenyan Xiong, Zhiyuan Liu, Yu Gu, and Ge Yu. MARVEL: Unlocking the multi-modal capability of dense retrieval via visual module plugin. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14608–14624, Bangkok, Thailand, 2024. Association for Computational Linguistics. 1, 2

[76] Fengbin Zhu, Wenqiang Lei, Fuli Feng, Chao Wang, Haozhou Zhang, and Tat-Seng Chua. Towards complex document understanding by discrete reasoning. In *MM '22: The 30th ACM International Conference on Multimedia, Lisboa, Portugal, October 10 - 14, 2022*, pages 4857–4866. ACM, 2022. 3