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

Supplementary

1. UMRB Details

Table 1 summarizes all UMRB tasks along with their statistics. Table 9 provides examples of different task types. Below is a brief description of each dataset included in the UMRB.

1.1. Single-Modal Tasks

WebQA [1] This dataset is derived from Wikipedia. In the T→T setup, both the query and candidate are text. The objective is to find a Wikipedia paragraph that answers the question. We have used 2,455 samples as the test set.

Nights [5] This dataset contains human judgments on the similarity of various image pairs, where both the query and candidate are images. The task is to identify an image that resembles the provided query image. We included 2,120 samples in our UMRB.

ArguAna, ClimateFEVER, CQADupstack, DBPedia, FEVER, FiQA2018, HotpotQA, MSMARCO, NFCorpus, NQ, Quora, SCIDOCS, SciFact, Touche2020 and TRECCOVID For these datasets, we use the processed versions from BEIR [18].

1.2. Cross-Modal Tasks

VisualNews [11] This dataset focuses on the news domain and consists of pairs of news headlines and associated images. In UMRB, this dataset can be transformed into two tasks: retrieving the corresponding image based on the news headline (T→I) and retrieving the corresponding news headline based on the image (I→T). We utilized 19,995 and 20,000 samples to construct the test set.

Fashion200k [6] This dataset includes pairs of images and product descriptions. In total, we have 1,719 instances for the task T→I and 4,889 instances for the task I→T for evaluation.

MSCOCO [9] This dataset is a well-known image caption dataset. Similar to VisualNews, it is converted into two tasks: "I→T", which retrieves the caption given an image and "T→I", which retrieves the image given a caption.

Flickr30k[17] This dataset consists of images paired with detailed textual descriptions. We have a total of 1,000 instances for the I→T task and 5,000 instances for the T→I task available for evaluation.

TAT-DQA, ArxivQA, DocVQA, InfoVQA, Shift Project, Artificial Intelligence, Government Reports, Healthcare Industry, Energy, TabFQuad These datasets constitute the retrieval task of T→VD. Their queries are standard questions, and the candidates are document screenshots. For these datasets, we used the processed versions from ViDoRe [4].

1.3. Fused-Modal Tasks

WebQA [1] Similar to WebQA in the Single-Modal setting, this dataset is also derived from Wikipedia, but in the T→IT setup, the candidates consist of images and text. The task is to find a Wikipedia paragraph with accompanying text and images to answer a specific question. There are 2,511 samples in the evaluation set.

EDIS [12] This dataset involves the cross-modal image search within the news domain. The queries are texts containing entities and events, with candidates consisting of news images and their accompanying headlines. The task requires the model to comprehend both entities and events from the text queries and retrieve the corresponding image and headline.

OVEN [8] The dataset is sourced from Wikipedia, where a query consists of an image and a question related to the image. The candidates are the Wikipedia title along with the first 100 tokens of its summary. If the associated Wikipedia content includes images, it constitutes an IT→IT task; otherwise, it forms an IT→T task. In the evaluation, we have 14,741 samples for the IT→IT task and 50,004 samples for the IT→T task.

Name	Type	Categ.	Eval Samples	Candidates Nums	Eval Query avg. chars	Eval Candidate avg. chars	In partial
ArguAna	Single-Modal	T→T	10,080	1,406	192.98	166.80	True
Climate-FEVER	Single-Modal	T→T	1,535	5,416,593	20.13	84.76	False
CQADupStack	Single-Modal	T→T	13,145	457,199	8.59	129.09	False
DBPedia	Single-Modal	T→T	400	4,635,922	5.39	49.68	False
FEVER	Single-Modal	T→T	6,666	5,416,568	8.13	84.76	False
FiQA2018	Single-Modal	T→T	648	57,638	10.77	132.32	False
HotpotQA	Single-Modal	T→T	7,405	5,233,329	17.61	46.30	False
MSMARCO	Single-Modal	T→T	6,980	8,841,823	5.96	55.98	False
NFCorpus	Single-Modal	T→T	323	3,633	3.30	232.26	True
NQ	Single-Modal	T→T	3,452	2,681,468	9.16	78.88	False
Quora	Single-Modal	T→T	10,000	522,931	9.53	11.44	True
SCIDOCS	Single-Modal	T→T	1,000	25,657	9.38	176.19	True
SciFact	Single-Modal	T→T	300	5,183	12.37	213.63	False
Touche2020	Single-Modal	T→T	49	382,545	6.55	292.37	False
TRECCOVID	Single-Modal	T→T	50	171,332	10.60	160.77	True
WebQA	Single-Modal	T→T	2,455	544,457	18.58	37.67	False
Nights	Single-Modal	I→I	2,120	40,038	-	-	True
VisualNews	Cross-Modal	T→I	19,995	542,246	18.78	-	False
Fashion200k	Cross-Modal	T→I	1,719	201,824	4.89	-	False
MSCOCO	Cross-Modal	T→I	24,809	5,000	10.43	-	True
Flickr30k	Cross-Modal	T→I	5,000	1,000	12.33	-	True
TAT-DQA	Cross-Modal	T→VD	1,646	277	12.44	-	False
ArxivQA	Cross-Modal	T→VD	500	500	17.12	-	False
DocVQA	Cross-Modal	T→VD	451	500	8.23	-	True
InfoVQA	Cross-Modal	T→VD	494	500	11.29	-	False
Shift Project	Cross-Modal	T→VD	100	1,000	16.01	-	True
Artificial Intelligence	Cross-Modal	T→VD	100	968	12.3	-	False
Government Reports	Cross-Modal	T→VD	100	972	12.62	-	False
Healthcare Industry	Cross-Modal	T→VD	100	965	12.56	-	False
Energy	Cross-Modal	T→VD	100	977	13.49	-	False
TabFQuad	Cross-Modal	T→VD	280	70	16.49	-	False
VisualNews	Cross-Modal	I→T	20,000	537,568	-	18.53	False
Fashion200k	Cross-Modal	I→T	4,889	61,707	-	4.95	False
MSCOCO	Cross-Modal	I→T	5,000	24,809	-	10.43	True
Flickr30k	Cross-Modal	I→T	1,000	5,000	-	12.33	True
WebQA	Fused-Modal	T→IT	2,511	403,196	16.43	12.83	False
EDIS	Fused-Modal	T→IT	3,241	1,047,067	20.07	15.53	False
OVEN	Fused-Modal	IT→T	50,004	676,667	6.52	82.13	False
INFOSEEK	Fused-Modal	IT→T	11,323	611,651	8.76	91.49	False
ReMuQ	Fused-Modal	IT→T	3,609	138,794	13.82	34.26	True
OKVQA	Fused-Modal	IT→T	5,046	114,516	8.09	102.55	True
LLaVA	Fused-Modal	IT→T	5,120	5,994	10.70	90.65	True
FashionIQ	Fused-Modal	IT→I	6,003	74,381	11.70	-	True
CIRR	Fused-Modal	IT→I	4,170	21,551	11.01	-	True
OVEN	Fused-Modal	IT→IT	14,741	335,135	5.91	94.76	True
EVQA	Fused-Modal	IT→IT	3,743	68,313	9.38	211.12	False
INFOSEEK	Fused-Modal	IT→IT	17,593	481,782	7.94	96.00	False

Table 1. Tasks in UMRB. We counted the number of datasets under each task type and the number of evaluation instances in the dataset, the size of the candidate set, and the average length of the text.

INFOSEEK [2] This dataset is similar to OVEN, with queries consisting of images alongside text questions. The candidates are Wikipedia snippets of 100 tokens containing the exact answers. This dataset also encompasses two tasks: for the IT→IT and IT→T tasks, we used 17,593 and 11,323 samples, respectively.

ReMuQ [14] The dataset is augmented from the WebQA questions by adding images to create new multimodal queries along with a large text corpus. For evaluation, we used 3,609 instances from this dataset.

OKVQA [15] This dataset includes visual questions that require external knowledge to answer. It is structured as an IT→T retrieval task, where queries consist of visual questions containing images and text, with candidates being external knowledge sources that can assist in answering the questions.

LLaVA [10] This dataset contains high-quality conversations about an image generated by GPT-3.5, involving exchanges between a human and an AI assistant. The queries comprise questions and instructions sent by humans to the AI assistant, which include both images and text, while the candidates are the AI assistant's replies. We utilized 5,120 samples from this dataset in the UMRB evaluation.

FashionIQ [21] This dataset features images of fashion products along with crowd-sourced descriptions that highlight the differences between these products. Each query consists of an image and a modification sentence that describes changes to the given image, with the retrieval target being the specified image. In the UMRB evaluation, we used 6,003 samples from this dataset.

CIRR [13] Similar to FashionIQ, CIRR can also be used for composed image retrieval. It involves pairs of real-life reference and target images in each test case, along with a modification sentence detailing the differences between the two images. For the UMRB evaluation, we utilized 4,170 samples from this dataset.

EVQA [16] This dataset is akin to INFOSEEK, with the key distinction being that the retrieval target of EVQA is a complete Wikipedia paragraph with a maximum length of several thousand tokens. We used 3,743 samples for evaluation, eliminating multi-hop issues present in the original test set. We selected Wikipedia paragraphs from the original dataset as candidates and supplemented them with images. Images native to each paragraph were included when available; otherwise, the first image from the article was utilized due to its typically representative nature.

1.4. UMRB-Partial

The full UMRB dataset consists of 47 subtasks, approximately 200,000 evaluation instances, and 40 million candidates, resulting in a significant overhead when testing the model. During our experiments with the GME-7B model, a full evaluation required approximately 400 A100*80G GPU hours. To facilitate development and verification, we created a smaller benchmark by condensing the complete UMRB, which we refer to as **UMRB-Partial**. Column 8 of Table 1 indicates whether a dataset is included in **UMRB-Partial**. Testing the GME-7B model on UMRB-Partial reduced the evaluation time from 400 A100*80G GPU hours to 80 A100*80G GPU hours.

2. Results Details

In this section, we present the detailed scores achieved by our GME and the baseline models on various tasks. Additionally, we provide results from other benchmarks, including BEIR, M-BEIR, and ViDoRe.

2.1. Detailed Results on UMRB

Table 2 presents the detailed evaluation results of the baseline systems alongside our GME on UMRB tasks. 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). The larger model, *i.e.* `GME-Qwen2-VL-7B`, further enhances this performance. In addition, focusing on specific scores on different datasets, our GME achieves state-of-the-art performance on each dataset except the Nights dataset. VISTA and CLIP-SF scored highly on the Nights dataset, likely due to their use of independent image and text encoders for cross-modal retrieval. In the I→I task, these models relied solely on the image encoder for encoding without cross-modal alignment, which may explain their superior performance on the Nights dataset.

2.2. Detailed Results on UMRB-Partial

Figure 3 of main paper illustrates our exploration of the training data, as discussed in Section 4.2, with specific results presented in Table 3. This table details the scores of our models trained on six data types: T→T, I→I, T→VD, T→I, IT→IT, and Mix across various tasks. We find that the model trained on mixed data performs the best.

2.3. Detailed Results on BEIR

BEIR is a heterogeneous benchmark containing diverse text IR tasks. We utilize BEIR to compare the performance of our GME with other text embedders on T→T tasks. Table 4 presents the detailed evaluation nDCG@10 scores for pure text embedders and multimodal embedders on T→T tasks. Except for our GME, other multimodal embedders do not match the performance of pure text embedders on text retrieval tasks, including those like `E5-V` that are fine-tuned exclusively on text data.

Naturally, pure text embedding models of the same model size still outperform multimodal embedding models in pure text retrieval tasks. For example, the score of the

Type	Task	Dataset	VISTA	CLIP-SF	One-Peace	DSE	E5-V	GME-2B	GME-7B
Single-Modal	T→T	ArguAna	63.61	52.45	32.93	53.46	54.28	63.18	72.11
		Climate-FEVER	31.17	20.00	20.27	19.79	21.64	41.08	48.36
		CQADupStack	42.35	30.61	41.32	36.51	41.69	39.06	42.16
		DBPedia	40.77	26.37	32.43	40.75	38.78	41.00	46.30
		FEVER	86.29	50.58	51.91	80.12	78.99	92.06	93.81
		FiQA2018	40.65	22.14	36.79	36.2	45.41	43.8	63.23
		HotpotQA	72.6	41.33	46.51	70.79	60.88	65.3	68.18
		MSMARCO	41.35	22.15	36.55	37.73	41.23	40.61	42.93
		NFCorpus	37.39	27.05	31.6	32.82	36.97	38.84	36.95
		NQ	54.15	25.45	42.87	52.97	51.58	54.52	56.08
		Quora	88.90	81.63	87.46	85.84	87.6	88.12	89.67
		SCIDOCS	21.73	14.75	21.64	15.66	22.36	22.94	26.35
		SciFact	74.04	55.98	64.51	68.97	72.75	74.19	82.43
		Touche2020	25.7	17.47	16.90	14.50	21.61	26.57	22.55
		TRECCOVID	77.90	63.61	69.28	52.98	72.85	71.73	77.49
		WebQA	83.80	84.44	63.67	83.95	89.94	94.34	94.34
	I→I	Nights	24.43	31.42	31.27	27.36	27.92	30.61	30.57
Cross-Modal	T→I	VisualNews	5.77	42.80	48.95	14.12	29.46	39.20	46.27
		Fashion200k	3.08	18.38	32.34	3.08	3.78	23.50	27.64
		MSCOCO	47.97	80.75	71.45	74.62	52.38	76.22	79.77
		Flickr30k	74.68	94.28	92.78	94.42	77.38	94.5	97.38
	T→VD	TAT-DQA	2.05	5.49	14.44	49.01	9.08	57.88	64.06
		ArxivQA	10.30	24.10	43.94	78.17	41.16	81.41	82.55
		DocVQA	8.01	11.80	23.48	45.83	24.37	46.86	49.34
		InfoVQA	30.02	48.78	59.97	82.06	49.5	84.97	88.79
		Shift Project	3.26	6.06	17.02	69.84	13.16	77.94	83.5
		Artificial Intelligence	7.34	28.64	45.41	96.88	46.18	95.75	98.02
		Government Reports	6.90	34.67	55.98	92.04	53.05	92.05	94.05
		Healthcare Industry	9.39	32.64	59.55	96.35	59.61	96.08	97.29
		Energy	11.05	27.19	53.21	92.62	56.77	89.17	93.09
		TabFQuad	13.08	21.53	57.05	79.29	58.22	91.79	94.92
	I→T	VisualNews	2.79	42.67	47.27	8.74	29.54	38.21	47.16
		Fashion200k	4.72	18.10	30.89	3.91	4.62	26.61	31.05
		MSCOCO	48.92	91.94	85.6	82.06	86.4	85.18	85.92
		Flickr30k	68.50	99.11	98.60	97.11	89.62	99.00	98.9
Fused-Modal	T→IT	WebQA	54.84	78.42	32.42	66.99	49.62	82.24	84.11
		EDIS	36.78	54.09	53.01	41.26	49.62	68.10	77.40
	IT→T	OVEN	22.32	45.98	23.69	0.38	14.4	59.67	64.13
		INFOSEEK	18.53	27.58	20.05	3.06	12.69	39.22	34.67
		ReMuQ	76.20	83.71	26.41	94.60	52.15	96.73	95.48
		OKVQA	17.14	17.44	9.67	13.28	16.71	30.08	32.61
		LLaVA	72.81	91.91	51.64	53.18	77.48	98.93	98.18
	IT→I	FashionIQ	3.28	24.54	2.93	9.81	3.73	26.34	29.89
		CIRR	14.65	45.25	10.53	36.52	13.19	47.70	51.79
	IT→IT	OVEN	27.77	68.83	30.56	0.39	54.46	78.96	83.05
		EVQA	28.75	40.08	16.64	15.34	26.39	77.32	79.88
		INFOSEEK	22.27	49.05	23.32	5.96	39.69	41.14	31.58
Avg.			37.32	43.66	42.01	50.04	42.52	63.42	65.87

Table 2. The detailed results of the baselines and our GME on UMRB. Following previous works [4, 18, 20], 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.

Type	Task	Dataset	T→T	I→I	T→VD	T→I	IT→IT	Mix
Single-Modal	T→T	Arguan	56.25	43.51	56.73	33.53	53.22	56.22
		NFCorpus	35.23	28.89	33.23	33.18	30.48	35.76
		Quora	87.82	74.37	86.32	86.43	85.2	87.4
		SCIDOCS	19.07	11.82	17.51	17.2	16.93	19.88
		TRECCOVID	75.57	47.89	50.89	72.37	58.92	76.38
	I→I	Nights	27.97	28.11	24.9	28.53	26.04	30.85
Cross-Modal	T→I	MSCOCO	59.7	59.41	63.67	76.91	44.97	75.3
		Flickr30k	83.92	65.52	87.32	93.18	74.52	93.06
	T→VD	DocVQA	35.8	24.24	48.38	40.58	28.05	45.62
		Shift Project	57.86	45.47	77.08	50.36	53.12	74.84
	I→T	MSCOCO	74.72	63.82	80.46	84.64	70.48	84.24
		Flickr30k	94.1	82.5	96.3	97.2	90.1	97.5
Fused-Modal	IT→T	LLaVA	92.75	89.05	86.02	89.24	88.73	95.02
		ReMuQ	89.61	85.47	76.45	85.12	86.73	89.75
		OKVQA	24.55	16.6	15.78	16.92	18.57	20.23
	IT→I	FashionIQ	5.53	4.2	5.43	8.86	11.08	11.89
		CIRR	17.24	15.04	15.42	17.5	25.71	29.86
	IT→IT	OVEN	59.81	38.42	57.31	56.69	65.08	63.04
Avg.			55.42	45.80	54.50	54.91	51.55	60.38

Table 3. Performance of models trained on different data types on UMRB-partial. We present NDCG@10 scores for T→T tasks. For T→VD tasks, we provide NDCG@5 scores. For the FashionIQ dataset, we report Recall@10 scores, while for all other datasets, we report Recall@5 scores.

`gte-Qwen2-7B-instruct` model is 60.25, while the `GME-Qwen2-VL-7B` model, with the same model scale, scores 55.63. Although both models share the same text LLM, incorporating or extending multimodal capabilities leads to additional compromises in pure text performance. Minimizing this kind of loss remains an important research question.

2.4. Detailed Results on M-BEIR

M-BEIR, a multimodal benchmark for IR, serves as a comprehensive large-scale retrieval benchmark designed to evaluate multimodal retrieval models. As shown in Table 5, we report Recall@10 scores for the Fashion200K and FashionIQ datasets, while Recall@5 scores are provided for all other datasets. In M-BEIR, our GME continues to demonstrate state-of-the-art performance, underscoring the effectiveness of our approach.

2.5. Detailed Results on ViDoRe

ViDoRe represents the Visual Document Retrieval Benchmark, encompassing various page-level screenshot retrieval tasks. This benchmark includes the T→VD tasks within our UMRB. Table 6 presents the detailed nDCG@5 scores for our GME and other models. Our smaller model, *i.e.* `GME-Qwen2-VL-2B`, surpasses the previous state-of-the-art model (ColPali), which was exclusively trained on this dataset for this specific task. The larger model, *i.e.* `GME-Qwen2-VL-7B`, further improves upon this performance.

3. Experiment Details

3.1. Training Details

Our GME models (both 2B and 7B) are initialized using the Qwen2-VL [19] model series. We employ the `transformers` library for training in BF16 precision. The training utilizes Low-Rank Adaptation (LoRA) [7] with a rank of 8. We apply a decoupled AdamW optimizer with a learning rate and a weight decay of 1e-4. Additional hyperparameters are detailed in Table 7.

In our contrastive learning approach, we develop dense multimodal representation models (embedders) that utilize the [EOS] hidden state as the embedding of the input. The temperature for contrastive learning is set to 0.03. For each query, we include one positive candidate along with eight hard negative candidates.

3.2. Instructions

The complete UMRB consists of 47 tasks, each with distinct retrieval candidates and varying domains. Even within the same dataset, retrieval candidates can differ based on task types. For example, the WebQA dataset aims to retrieve textual candidates for T→T tasks, which is different from retrieving a combination of image and text candidates for T→IT tasks.

We have designed specific instructions tailored for each task to guide the model in effectively completing the retrieval process. The detailed instructions are provided in Table 8.

4. Fused-Modal Data Synthesis Details

We utilize doc2query to synthesize data. However, our goal is to generate fused-modal candidate-to-query relevance data rather than single-modality, text-based relevance pairs.

4.1. Prompts

Step 1: In the first step of data synthesis, we prompt the large language model (LLM) to generate a natural question and answer based on a selected passage. The specific prompt is illustrated in Figure 1. This process leverages in-context learning (ICL) to guide the LLM in producing outputs that align with our requirements.

Step 2: In step 2, we provide the LLM with the passage and the natural question generated in step 1. The LLM is then prompted to extract the main entity from the question and refactor the question accordingly. Figure 2 presents the prompt used in this step. In subsequent steps, the extracted entity will be replaced by the corresponding image, which, when combined with the reconstructed question, will form a fused-modal query.

BEIR	Avg.	Argu-Ana	Cli-mate-Fever	CQA-Dup-Stack	DB-Pedia	Fever	FiQA	Hotpot-QA	MS MAR-CO	NF-Corpus	NQ	Quora	Sci-docs	Sci-fact	Touche-2020	Trec-Covid
Text Embedder																
gte-Qwen2-7B-instruct	60.25	64.27	45.88	46.43	52.42	95.11	62.03	73.08	45.98	40.6	67	90.09	28.91	79.06	30.57	82.26
NV-Embed-v1	59.36	68.2	34.72	50.51	48.29	87.77	63.1	79.92	46.49	38.04	71.22	89.21	20.19	78.43	28.38	85.88
gte-Qwen2-1.5B-instruct	58.29	69.72	42.91	44.76	48.69	91.57	54.7	68.95	43.36	39.34	64	89.64	24.98	78.44	27.89	85.38
voyage-large-2-instruct	58.28	64.06	32.65	46.6	46.03	91.47	59.76	70.86	40.6	40.32	65.92	87.4	24.32	79.99	39.16	85.07
neural-embedding-v1	58.12	67.21	32.3	49.11	48.05	89.46	58.94	78.87	42	42.6	68.36	89.02	27.69	78.82	24.06	75.33
GritLM-7B	57.41	63.24	30.91	49.42	46.6	82.74	59.95	79.4	41.96	40.89	70.3	89.47	24.41	79.17	27.93	74.8
e5-mistral-7b-instruct	56.89	61.88	38.35	42.97	48.89	87.84	56.59	75.72	43.06	38.62	63.53	89.61	16.3	76.41	26.39	87.25
google-gecko	55.7	62.18	33.21	48.89	47.12	86.96	59.24	71.33	32.58	40.33	61.28	88.18	20.34	75.42	25.86	82.62
text-embedding-3-large	55.44	58.05	30.27	47.54	44.76	87.94	55	71.58	40.24	42.07	61.27	89.05	23.11	77.77	23.35	79.56
gte-en-large-v1.5	57.91	72.11	48.36	42.16	46.3	93.81	63.23	68.18	42.93	36.95	56.08	89.67	26.35	82.43	22.55	77.49
gte-en-base-v1.5	54.09	63.49	40.36	39.52	39.9	94.81	48.65	67.75	42.62	35.88	52.96	88.42	21.92	76.77	25.22	73.13
Multimodal Embedder																
VISTA	53.24	63.61	31.17	42.35	40.77	86.29	40.65	72.6	41.35	37.39	54.15	88.9	21.73	74.04	25.7	77.9
CLIP-SF	36.77	52.45	20	30.61	26.37	50.58	22.14	41.33	22.15	27.05	25.45	81.63	14.75	55.98	17.47	63.60
One-Peace	42.19	32.93	20.27	41.32	32.43	51.91	36.79	46.51	36.55	31.6	42.87	87.46	21.64	64.51	16.9	69.28
DSE	46.60	53.46	19.79	36.51	40.75	80.12	36.2	70.79	37.73	32.82	52.97	85.84	15.66	68.97	14.50	52.98
E5-V	49.91	54.28	21.64	41.69	38.78	78.99	45.41	60.88	41.23	36.97	51.58	87.6	22.36	72.75	21.61	72.85
GME-Qwen2-VL-2B	53.31	61.52	42.3	38.13	46.31	92.6	45.3	72.93	40.88	37.2	60.01	87.24	23.17	63.82	29.06	59.24
GME-Qwen2-VL-7B	55.68	64.60	45.38	41.66	50.78	94.27	57.14	79.21	42.38	38.40	67.74	88.05	27.38	62.31	23.26	52.6

Table 4. BEIR benchmark [18] nDCG@10 scores. We include top models from MTEB Retrieval English leaderboard.

MBEIR	Avg.	$q_t \to c_i$			$q_t \to c_t$	$q_t \to (c_i, c_t)$		$q_i \to c_t$			$q_i \to c_i$	$(q_i, q_t) \to c_t$		$(q_i, q_t) \to c_i$		$(q_i, q_t) \to (c_i, c_t)$	
		Visual-News	MS-COCO	Fashion-200K	Web-QA	EDIS	Web-QA	Visual-News	MS-COCO	Fashion-200K	NIGHTS	OVEN	Info-Seek	Fashion-IQ	CIRR	OVEN	Info-Seek
CLIP	32.5	43.3	61.1	6.6	36.2	43.3	45.1	41.3	79.0	7.7	26.1	24.2	20.5	7.0	13.2	38.8	26.4
SigLIP	37.2	30.1	75.7	36.5	39.8	27.0	43.5	30.8	88.2	34.2	28.9	29.7	25.1	14.4	22.7	41.7	27.4
BLIP	26.8	16.4	74.4	15.9	44.9	26.8	20.3	17.2	83.2	19.9	27.4	16.1	10.2	2.3	10.6	27.4	16.6
BLIP2	24.8	16.7	63.8	14.0	38.6	26.9	24.5	15.0	80.0	14.2	25.4	12.2	5.5	4.4	11.8	27.3	15.8
VISTA	26.37	5.77	47.97	3.08	83.80	36.78	54.84	2.79	48.92	4.72	24.43	22.32	18.53	3.28	14.65	27.77	22.27
CLIP-SF	50.26	42.80	80.75	18.38	84.44	54.09	78.42	42.67	91.94	18.10	31.42	45.98	27.58	24.53	45.25	68.83	49.05
One-Peace	38.00	48.95	71.45	32.34	63.67	53.01	32.42	47.27	85.60	30.89	31.27	23.69	20.05	2.93	10.53	30.56	23.32
DSE	28.89	14.12	74.62	3.08	83.95	41.26	66.99	8.74	82.06	3.91	27.36	0.38	3.06	9.81	36.52	0.39	5.96
E5-V	35.09	29.46	52.38	3.78	89.94	49.62	49.62	29.54	86.40	4.62	27.92	14.40	12.69	3.73	13.19	54.46	39.69
GME-Qwen2-VL-2B	53.54	38.85	71.82	25.83	95.19	70.32	83.15	38.32	84.12	27.57	29.86	58.17	39.06	27.5	46.83	75.98	44.21
GME-Qwen2-VL-7B	54.50	46.54	75.14	31.82	95.85	77.29	84.59	45.54	64.90	34.20	31.89	63.41	43.14	31.43	53.69	80.30	58.80

Table 5. Results of M-BEIR benchmark [20]. For the Fashion200K and FashionIQ datasets, we report Recall@10 scores, while for all other datasets, we report Recall@5 scores.

Step 3: In step 3, we replace the entity with an image, which can be sourced in two ways. The first method involves prompting the LLM to generate a caption for the entity based on the provided entity and passage, after which the caption is fed into FLUX to generate images. The second method retrieves the entity by utilizing the Google Image Retrieval API. Figure 3 illustrates the caption generation prompt for this step.

4.2. Filter

Two filtering methods are implemented to ensure the quality of the synthesized data. First, a text retrieval model is utilized to evaluate unreconstructed queries and their corresponding passages. We follow the framework of Prompttagator [3]; a query is deemed unqualified if the passage that generated it does not appear within the top 20 search results. Second, for images obtained through the Google Image Search API, we employ the CLIP model to assess image-caption relevance. Images with a relevance score below 0.2 are filtered out.

```
>> SYSTEM
You are a helpful assistant.

>> USER
Based on the given **Passage**, generate a query and answer. The
result should be returned in json format.
Here are some examples.
Example1:
**Passage**:
<passage>
**Output**:
{"query": "Is Heracleum mantegazzianum poisonous?", "answer": "yes"}

Now it's the **Passage** you have to deal with. Be careful to return
the result directly and not to generate other irrelevant information.
Remember the output should be returned in json format.
**Passage**:
<passage>
**Output**:
```

Figure 1. Fused-Modal Data Synthesis Step 1 Prompt.

Why is the threshold score set to 0.2? The relevance scores of all images searched via Google and the corresponding captions we have collected are presented in Figure 4. We select the median score of 0.2 to ensure image quality while also ensuring that most text queries have sufficient images to pair with.

	ArxivQ	DocQ	InfoQ	TabF	TATQ	Shift	AI	Energy	Gov.	Health.	Avg.
BM25$_{\text{Text + Captioning}}$	40.1	38.4	70.0	35.4	61.5	60.9	88.0	84.7	82.7	89.2	65.1
BGE-M3$_{\text{Text + Captioning}}$	35.7	32.9	71.9	69.1	43.8	73.1	88.8	83.3	80.4	91.3	67.0
Jina-CLIP	25.4	11.9	35.5	20.2	3.3	3.8	15.2	19.7	21.4	20.8	17.7
Nomic-vision	17.1	10.7	30.1	16.3	2.7	1.1	12.9	10.9	11.4	15.7	12.9
SigLIP (Vanilla)	43.2	30.3	64.1	58.1	26.2	18.7	62.5	65.7	66.1	79.1	51.4
ColPali	79.1	54.4	81.8	83.9	65.8	73.2	96.2	91.0	92.7	94.4	81.3
VISTA	10.3	8.01	30.02	13.08	2.05	3.26	7.14	11.05	6.9	9.39	10.12
CLIP-SF	24.1	11.8	48.78	21.53	5.49	6.06	28.64	27.19	34.67	32.64	24.09
One-Peace	43.94	23.48	59.97	57.05	13.44	17.02	45.41	53.21	55.98	59.5	42.9
DSE	78.17	45.83	82.06	79.29	49.01	69.84	96.89	92.62	92.04	96.35	78.21
E5-V	41.16	24.37	49.5	58.22	9.08	13.26	46.18	57.77	53.05	59.61	41.22
GME-Qwen2-VL-2B	83.91	54.57	91.11	94.61	71.05	94.29	99.02	93.15	97.89	98.89	87.84
GME-Qwen2-VL-7B	87.58	56.63	92.39	94.58	76.12	97.26	99.63	95.89	99.5	99.63	89.92

Table 6. Comprehensive evaluation of baseline models and our GME on *ViDoRe*. Results are presented using NDCG@5 metrics.

Hyper-param	GME-Qwen2-VL-2B	GME-Qwen2-VL-7B
Number of Params	2B	8.2B
Number of Layers	28	28
Hidden Size	1536	3584
FFN Inner Size	3072	
Number of Attention Heads	12	28
Vision Depth	32	
Vision Embed_dim	1280	
Vision Patch_size	14	
Temperature	0.03	
Learning Rate Decay	Linear	
Adam ϵ	1e-4	
Adam β_1	0.9	
Adam β_2	0.98	
Gradient Clipping	0.0	
Precision	PyTorch BF16 AMP	
Max Length	1800	1800
Batch Size	128	32
Warm-up Ratio	0.06	

Table 7. GME training hyper-parameters.

```
>> SYSTEM
You are a helpful assistant.

>> USER
Extract the entity corresponding to **Query** and **Passage**, and replace
the entity in query with general references, such as "this person, this
building, this animal, this river, this bridge....". The result is
returned in json format.
Here are some examples.
Example1:
**Query**:
Is Heracleum mantegazzianum poisonous?
**Passage**:
<passage>
**Output**:
{"entity":"Heracleum mantegazzianum","query": "Is this plant poisonous?"}

Now it's the **Query** and **Passage** you have to deal with. Be careful
to return the result directly and not to generate other irrelevant
information. Remember the output should be returned in json format.
**Query**:
<Query>
**Passage**:
<passage>
**Output**:
```

Figure 2. Fused-Modal Data Synthesis Step 2 Prompt.

4.3. Examples of synthetic data

Table 10 illustrates passages from 15 domains and the fused modal queries generated by applying the synthesis flow. "FLUX image" refers to images generated by the Vincennes diagram model FLUX.1-dev, whereas "Google image" indicates images from Google Image retrieval.

```
>> SYSTEM
You are a helpful assistant.

>> USER
Give an **Entity**, and a **Passage** introducing this entity. Generate a
concise **Description** of the appearance of the entity. The generated
description will be used to generate an image of the entity. The
description should be less than 25 words long.
Here are some examples.
Example1:
**Entity**:
Heracleum mantegazzianum
**Passage**:
<passage>
**Description**:
Heracleum mantegazzianum: a tall plant with large, compound leaves and
white, umbrella-like flower clusters.

Now it's the **Entity** and **Passage** you have to deal with. Be careful
to return the **Description** directly and not to generate other irrelevant
information. Remember the description should be less than 25 words long.
**Entity**:
<Entity>
**Passage**:
<passage>
**Description**:
```

Figure 3. Fused-Modal Data Synthesis Step 3 Prompt.

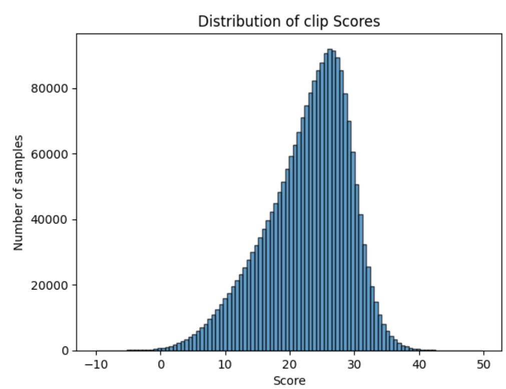


Figure 4. The distribution of relevance scores for all the images searched by Google and captions.

5. Limitations

In this work, we present a benchmark for training and testing Universal Multimodal Retrieval (UMR). To better accomplish this task, we explore strategies for adapting Mul-

Task	Dataset	Query Instruction
T→T	ArguAna	Given a claim, find documents that refute the claim.
	Climate-FEVER	Given a claim about climate change, retrieve documents that support orrefute the claim.
	CQADupStack	Given a question, retrieve detailed question descriptions from Stackexchange that are duplicates to the given question.
	DBPedia	Given a query, retrieve relevant entity descriptions from DBPedia.
	FEVER	Given a claim, retrieve documents that support or refute the claim.
	FiQA2018	Given a financial question, retrieve user replies that best answer the question.
	HotpotQA	Given a multi-hop question, retrieve documents that can help answer the question.
	MSMARCO	Given a web search query, retrieve relevant passages that answer the query.
	NFCorpus	Given a question, retrieve relevant documents that best answer the question.
	NQ	Given a question, retrieve Wikipedia passages that answer the question.
	Quora	Given a question, retrieve questions that are semantically equivalentto the given question.
	SCIDOCS	Given a scientific paper title, retrieve paper abstracts that are cited bythe given paper.
	SciFact	Given a scientific claim, retrieve documents that support or refute theclaim.
	Touche2020	Given a question, retrieve detailed and persuasive arguments that answer the question.
	TRECCOVID	Given a query on COVID-19, retrieve documents that answer the query.
	WebQA	Retrieve passages from Wikipedia that provide answers to the following question.
I→I	Nights	Find a day-to-day image that looks similar to the provided image.
T→I	VisualNews	Identify the news-related image in line with the described event.
	Fashion200k	Based on the following fashion description, retrieve the best matching image.
	MSCOCO	Identify the image showcasing the described everyday scene.
	Flickr30k	Find an image that matches the given caption.
T→VD	TAT-DQA ArxivQA DocVQA InfoVQA Shift Project Artificial Intelligence Government Reports Healthcare Industry Energy TabFQuad	Find a screenshot that relevant to the user's question.
I→T	VisualNews	Find a caption for the news in the given photo.
	Fashion200k	Find a product description for the fashion item in the image.
	MSCOCO	Find an image caption describing the following everyday image.
	Flickr30k	Find an image caption describing the following image.
T→IT	WebQA	Find a Wikipedia image that answers this question.
	EDIS	Identify the news photo for the given caption.
IT→T	OVEN INFOSEEK	Retrieve a Wikipedia paragraph that provides an answer to the given query about the image.
	ReMuQ	Retrieve a fact-based paragraph that provides an answer to the given query about the image.
	OKVQA	Retrieve documents that provide an answer to the question alongside the image.
	LLaVA	Provide a specific decription of the image along with the following question.
IT→I	FashionIQ	Find a fashion image that aligns with the reference image and style note.
	CIRR	Retrieve a day-to-day image that aligns with the modification instructions of the provided image.
IT→IT	OVEN INFOSEEK	Retrieve a Wikipedia image-description pair that provides evidence for the question of this image.
	EVQA	Obtain illustrated documents that correspond to the inquiry alongside the provided image.

Table 8. The instructions for different tasks, we only use the instructions for query encoding.

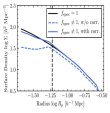
Type	Task	Query Text	Query Image	Target Text	Target Image
Single-Modal	T→T	where is whitemarsh island?	-	Whitemarsh Island, Georgia Whitemarsh Island, Georgia. Whitemarsh Island (pronounced WIT-marsh) is a census-designated place (CDP) in Chatham County, Georgia, United States. The population was 6,792 at the 2010 census. It is part of the Savannah Metropolitan Statistical Area. The communities of Whitemarsh Island are a relatively affluent suburb of Savannah.	-
	I→I	-		-	
Cross-Modal	T→I	Multicolor boutique amy black leather look biker jacket.	-	-	
	T→VD	Based on the graph, what is the impact of correcting for fspec not equal to 1 on the surface density trend?	-	-	
	I→T		-	Indian National Congress Vice President Rahul Gandhi addresses the special plenary session of Confederation of Indian Industr in New Delhi on April 4 2013.	-
Fused-Modal	T→IT	Does a Minnetonka Rhododendron flower have petals in a cup shape?	-	2020-05-08 15 17 05 Minnetonka Rhododendron flower along Tranquility Court in the Franklin Farm section of Oak Hill, Fairfax County, Virginia Minnetonka Rhododendron flower along Tranquility Court in the Franklin Farm section of Oak Hill, Fairfax County, Virginia.	
	IT→T	What is this plant named after?		Kalmia. Kalmia is a genus of about ten species of evergreen shrubs from 0.2–5 m tall, in the family Ericaceae. They are native to North America ... saw it during his travels in Carolina, and after his return to England in.	-
	IT→I	Is shiny and silver with shorter sleeves and fit and flare.		-	
	IT→IT	Is this plant poisonous?		Heracleum mantegazzianum, commonly known as giant hogweed, is a monocarpic perennial herbaceous plant in the carrot family Apiaceae ...These serious reactions are due to the furanocoumarin derivatives in the leaves, roots, stems, flowers, and seeds of the plant. Consequently, it is considered to be a noxious weed in many jurisdictions.	

Table 9. Data examples in diffierent task type. Due to the limitations of the table, we have cropped the displayed text.

timodal Large Language Models (MLLMs) into UMR models, presenting GME, a powerful embedding model capable of retrieving candidates across different modalities. However, this work has its limitations, which are outlined below:

1. Single Image Limit In MLLMs, one image is converted into a very large number of visual tokens. In Qwen2-VL, we limit the number of visual tokens to 1024. Due to model training efficiency and a lack of relevant data, our queries and candidates in UMRB only retain a single image. Thus, performance on interleaved data (where multiple images and texts are mixed together) cannot be assessed.

2. Single Language Limit Although the backbone of our model, Qwen2-VL, supports multiple languages, we only utilized a single language, English, during the training and testing processes of our GME. Consequently, performance in other languages could not be evaluated.

References

[1] Yingshan Chang, Guihong Cao, Mridu Narang, Jianfeng Gao, Hisami Suzuki, and Yonatan Bisk. Webqa: Multihop

Domain	Candidate Image	Candidate Text	FLUX Image	Google Image	Query Text
animal		The golden poison frog is the most poisonous animal on the planet; these frogs produce deadly alkaloid batrachotoxins in their skin glands as a defense against predators. To become poisoned a predator generally must attempt to consume the frog, ... has modified sodium channels unaffected by batrachotoxin.			What is the primary defense mechanism of this animal?
architecture		Neoclassical buildings are characterized by their magnificence of scale, the prominent use of columns, the use of geometric forms and symmetry, ...Samriddhi Bhavan,...National library of India, Kolkata			What are some examples of this style in Indian public buildings?
artwork		"Finding Peace Under Pressure: A Close Look at the new Butterfly of Peace". The Houston Museum of Natural Science. Retrieved 2021-07-05."Aurora Butterfly of Peace on Display at Smithsonian". The Gemmological Association of Great Britain. Retrieved 2021-07-05.			Where was this display shown?
currency		The euro was founded on 1 January 1999, when it became the currency of over 300 million people in Europe. For the first three years of its existence it was an... Slovenia joined the Eurozone in 2007, Cyprus and Malta in 2008, Slovakia in 2009, Estonia in 2011 and Latvia on 1 January 2014.			When did this currency become available?
entertainment		Thomas Middleditch as Richard Hendricks, a coder and founder/CEO of Pied Piper.T.J. Miller as Erlich Bachman (seasons 1–4), an Chris Diamantopoulos as Russ Hanneman ...a brash, loud and fiery billionaire investor who provides Pied Piper with their Series A.			Who is the CEO of this company in the TV series Silicon Valley?
food		An Italian beef sandwich features thin slices of seasoned roast beef, dripping with meat juices, on a dense, long Italian-style roll, believed to have originated in Chicago, where its history ... Despite the name, it is almost completely unknown in Italy.			What city is this sandwich believed to have originated in?
language		In the early 6th century BCE, the Neo-Babylonian Empire conquered the ancient Kingdom of Judah, destroying much of Jerusalem and exiling its population far to the East in Babylon. During ...details on Hebrew and Aramaic in the gospels.)			What languages were spoken in this region during the Roman period?
literature		The Adventures of Huckleberry Finn (1973), by Robert James Dixson – a simplified version Big River: The Adventures of Huckleberry Finn, a 1985 ... Classics imprint was released in November 2017.			What form of media was this book adapted into in 1985?
mythology		Throughout India, on contemporary poster art, Ganesha is portrayed with Sarasvati (goddess of culture and art) or Lakshmi (goddess of luck and prosperity) or both. Ganesha, Lakshmi and Sarswati ... to be the brother of Sarasvati and Lakshmi.			What is the relationship between this deity and Sarasvati in Maharashtra?
organization		During World War II, ARC operated the American Red Cross Clubmobile Service to provide servicemen with food, entertainment and "a connection home." In a ...During the Vietnam War 627 American women served in the ARC Supplemental Recreation Overseas Program. At the invitation			What service did this organization provide to boost soldier morale during the Vietnam War?
person		Runnels later re-emerged in 1998, under her real name, as the on-screen girlfriend of Val Venis. When Runnels claimed to be pregnant with Venis' baby, he dumped her... broke up by July, when Jacqueline Moore became frustrated with Runnels' infatuation with Meat.			Who did this person claim to be pregnant with in 1998?
pharmaceutical		DHA-paclitaxel (or Taxoprexin) is an investigational drug (from Protarga Inc) made by linking paclitaxel to docosahexaenoic acid (DHA), a fatty acid that is easily ...may be able to treat more types of cancer than Taxol has been able to treat.			What is the advantage of this drug over paclitaxel?
plant		The species was first described as Salpiglossis integrifolia by William Jackson Hooker in 1831. It was transferred to the genus Petunia as P. integrifolia by Hans Schinz and Albert Thellung... ranges, with P. inflata growing in more northern areas.			What was the original genus of this plant?
sport		The Columbia University Lions are the collective athletic teams and their members from Columbia University, an Ivy League institution in New York City, United States. The current director of athletics is Peter Pilling.			What is the name of the athletic teams from this university?
vehicle		A specialized Lexus LS 460 is used in a warehouse-sized driving simulator at Toyota's Higashifuji Technical Center in Shizuoka, Japan. This vehicle is mounted ... automotive safety features in a secure environment.			What is the purpose of this driving simulator at Toyota's Higashifuji Technical Center?

Table 10. Examples of synthetic data. Due to the limitations of the table, we have cropped the displayed text.

and multimodal QA. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16474–16483, 2022. 1

[2] 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. 2

[3] 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

[4] 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. 1, 4

[5] 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. 1

[6] 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. 1

[7] 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. 5

[8] 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. 1

[9] 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. 1

[10] 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

[11] 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. 1

[12] 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. 1

[13] 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

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

[15] 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

[16] 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

[17] 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. 1

[18] 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. 1, 4, 6

[19] 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. 5

[20] 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. 4, 6

[21] 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