

LamRA: Large Multimodal Model as Your Advanced Retrieval Assistant

Supplementary Material

A. Analysis of Pre-training Dataset Selection

As outlined in Section 3.3 of the main paper, the LamRA-Ret undergoes a two-stage training process to enhance retrieval performance incrementally. These stages consist of the language-only pre-training phase and the multimodal instruction-tuning phase. In this section, we conduct our pre-training experiments using various types of datasets, which can be categorized into the following groups: (i) datasets containing only image data, (ii) datasets comprising solely text data, (iii) datasets consisting of image-text pairs, and (iv) a combination of pure text data and image-text pair datasets. We utilize MSCOCO and Flickr30K as our evaluation benchmark for the pre-training stage.

Image-only Dataset. To construct suitable training data, we utilize the Region-based 100K dataset within UltraEdit [13], a region-based image editing dataset. The reference image is employed as the query, the reference image with appropriate data enhancements serves as the positive sample, and the edited image is used as the hard negative sample.

Language-only Dataset. We employ the NLI dataset introduced in [4], which comprises a series of triplets: (query text, positive text, and hard negative text). The dataset contains 275K samples. In our experiments, we observe that excluding the hard negative text yields improved results. Consequently, in the default experimental setting, we use only the query text and positive text while treating other samples within the same batch as negative samples.

Image-Text Pairs Dataset. We use the ShareGPT4V dataset [2], which comprises paired images and long texts, with an initial size of 100K samples. After excluding the COCO images, the final dataset size is reduced to 52K samples.

Combined Dataset. We construct our combined dataset using a combination of the language-only and image-text pairs data.

Results Analysis. The experimental results, as presented in Table 1, reveal the following observations: (i) Across different types of pertaining data, the retrieval performance consistently improves. (ii) Performance fluctuates depending on the data type, with language-only pertaining data yielding the best results. (iii) Simply combining different types of pretraining data can unexpectedly degrade performance. Based on these findings, we ultimately opt for the language-only pretraining approach.

Dataset Type	Image Retrieval						Text Retrieval					
	Flickr30K			COCO			Flickr30K			COCO		
	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10	R@1	R@5	R@10
No Training	52.4	78.9	86.0	26.4	50.9	62.9	54.7	80.8	88.3	32.3	55.3	66.6
Image-only	71.3	92.0	95.2	38.9	66.2	76.6	79.3	94.0	97.3	58.0	80.0	88.0
Language-only	80.8	95.3	97.2	53.3	77.0	84.8	91.2	99.0	99.5	67.5	87.1	92.6
Image-Text Pairs	72.5	90.7	94.4	46.0	71.1	79.9	88.4	97.5	99.0	65.3	85.4	91.3
Combined	74.1	92.5	95.8	47.2	71.8	80.4	83.9	96.0	98.4	55.1	78.9	86.3

Table 1. Performance across various Pre-training Dataset.

B. More Implementation Details

Supplement of Feature Extraction. As discussed in Section 3.2 of the main paper, the `<emb>` token is a newly introduced word added to the vocabulary, primarily serving as a placeholder. Essentially, we can use an arbitrary token to replace it, since the embedding we use is derived from the hidden state corresponding to the position preceding this token.

Supplement of Listwise Reranking Method. As discussed in Section 3.4 of the main paper, the pointwise reranking method assigns a score to each candidate based on the probability of the LMMs outputting YES for that candidate. Similarly, the listwise reranking method adopts a similar approach, assigning a reranking score to each candidate according to the probability of the LMM outputting a specific serial number.

Supplement of Baseline Methods. We utilize the official checkpoints of each baseline model for evaluation. We directly report the results of datasets that have already been evaluated in the corresponding papers. Note that, the Coca version of MagicLens [12] has not yet been open-sourced, therefore, we use its CLIP version for our experiments.

C. Details about M-BEIR Dataset

We present the details for the M-BEIR benchmark and the corresponding instruction in Table 2 and Table 4, respectively. It is important to note that the M-BEIR benchmark applies additional processing to the datasets it incorporates, which may result in differences from the standard evaluation of individual datasets. For instance, the candidate pool of the CIRR dataset in M-BEIR includes training data, which essentially increases the evaluation’s difficulty compared to the original CIRR dataset. For a more comprehensive understanding of these differences, we refer the readers to the original UniIR [11] paper.

Task	Dataset	Domain	# Train	# Dev	# Test	# Pool
$q^t \rightarrow c^i$	VisualNews	News	99K	20K	20K	542K
	MSCOCO	Misc.	100K	24.8K	24.8K	5K
	Fashion200K	Fashion	15K	1.7K	1.7K	201K
$q^t \rightarrow c^t$	WebQA	Wiki	16K	1.7K	2.4K	544K
$q^t \rightarrow (c^i, c^t)$	EDIS	News	26K	3.2K	3.2K	1M
	WebQA	Wiki	17K	1.7K	2.5K	403K
$q^i \rightarrow c^t$	VisualNews	News	100K	20K	20K	537K
	MSCOCO	Misc.	113K	5K	5K	25K
	Fashion200K	Fashion	15K	4.8K	4.8K	61K
$q^i \rightarrow c^i$	NIGHTS	Misc.	16K	2K	2K	40K
$(q^i, q^t) \rightarrow c^t$	OVEN	Wiki	150K	50K	50K	676K
	InfoSeek	Wiki	141K	11K	11K	611K
$(q^i, q^t) \rightarrow c^i$	FashionIQ	Fashion	16K	2K	6K	74K
	CIRR	Misc.	26K	2K	4K	21K
$(q^i, q^t) \rightarrow (c^i, c^t)$	OVEN	Wiki	157K	14.7K	14.7K	335K
	InfoSeek	Wiki	143K	17.6K	17.6K	481K
8 tasks	10 datasets	4 domains	1.1M	182K	190K	5.6M

Table 2. Summary of the M-BEIR benchmarks.

D. Details about Unseen Dataset

Here, we present the details of the Unseen Dataset in Table 3. Many of them are actually adapted from MSCOCO or FashionIQ, however, note that, their captions or query formats are significantly different. Therefore, we still treat these datasets as unseen datasets. For instance, the captions in Urban1K consist of extended captions generated by GPT-4V [10], while the query format of CIRCO combines a reference image with a relative caption. These differences create a substantial disparity compared to the original COCO dataset.

Dataset	Image Source	Task	Query Format	Candidate Format
ShareGPT4V	SA-1B	$q^t \rightarrow c^i$ $q^i \rightarrow c^t$	<long text> <image>	<image> <long text>
Urban-1K	MSCOCO	$q^t \rightarrow c^i$ $q^i \rightarrow c^t$	<long text> <image>	<image> <long text>
Flickr30K	Flickr	$q^t \rightarrow c^i$ $q^i \rightarrow c^t$	<short text> <image>	<image> <short text>
CIRCO	MSCOCO unlabeled set	$(q^i, q^t) \rightarrow c^i$	<image><relative caption>	<image>
GeneCIS	MSCOCO	$(q^i, q^t) \rightarrow c^i$	<image><relative caption>	<image>
Visual Dialog	MSCOCO	$q^{\text{dialog}} \rightarrow c^i$	<Q ₁ ><A ₁ >...<Q _j ><A _j >	<image>
Visual Storytelling	Flickr	$(q^i \oplus q^t) \rightarrow c^i$	<text ₁ ><image ₁ >...<text _j >	<image>
MT-FIQ	FashionIQ	$(q^i \oplus q^t) \rightarrow c^i$	<image ₁ ><relative caption ₁ >... <image _j ><relative caption _j >	<image>
CC-Neg	CC3M	ITM	<image>	<text>
Sugar-Crepe	MSCOCO	ITM	<image>	<text>

Table 3. Summary of the Unseen Dataset.

Task	Dataset	Instruction
$q^t \rightarrow c^i$	VisualNews	Identify the news-related image in line with the described event. Display an image that best captures the following caption from the news. Based on the caption, provide the most fitting image for the news story. I want you to retrieve an image of this news caption.
	MSCOCO	Find me an everyday image that matches the given caption. Identify the image showcasing the described everyday scene. I want to retrieve an image of this daily life description. Show me an image that best captures the following common scene description.
	Fashion200K	Based on the following fashion description, retrieve the best matching image. Match the provided description to the correct fashion item photo. Identify the fashion image that aligns with the described product. You need to identify the image that corresponds to the fashion product description provided.
$q^t \rightarrow c^t$	WebQA	Retrieve passages from Wikipedia that provide answers to the following question. You have to find a Wikipedia paragraph that provides the answer to the question. I want to find an answer to the question. Can you find some snippets that provide evidence from Wikipedia? I'm looking for a Wikipedia snippet that answers this question.
$q^t \rightarrow (c^i, c^t)$	EDIS	Find a news image that matches the provided caption. Identify the news photo for the given caption. Can you pair this news caption with the right image? I'm looking for an image that aligns with this news caption.
	WebQA	Find a Wikipedia image that answers this question. Provide with me an image from Wikipedia to answer this question. I want to know the answer to this question. Please find the related Wikipedia image for me. You need to retrieve an evidence image from Wikipedia to address this question.
$q^i \rightarrow c^t$	VisualNews	Find a caption for the news in the given photo. Based on the shown image, retrieve an appropriate news caption. Provide a news-related caption for the displayed image. I want to know the caption for this news image.
	MSCOCO	Find an image caption describing the following everyday image. Retrieve the caption for the displayed day-to-day image. Can you find a caption talking about this daily life image? I want to locate the caption that best describes this everyday scene image.
	Fashion200K	Find a product description for the fashion item in the image. Based on the displayed image, retrieve the corresponding fashion description. Can you retrieve the description for the fashion item in the image? I want to find a matching description for the fashion item in this image.
$q^i \rightarrow c^i$	NIGHTS	Find a day-to-day image that looks similar to the provided image. Which everyday image is the most similar to the reference image? Find a daily life image that is identical to the given one. You need to identify the common scene image that aligns most with this reference image.
$(q^i, q^t) \rightarrow c^t$	OVEN	Retrieve a Wikipedia paragraph that provides an answer to the given query about the image. Determine the Wikipedia snippet that identifies the visual entity in the image. I want to find a paragraph from Wikipedia that answers my question about this image. You have to find a Wikipedia segment that identifies this image's subject.
	InfoSeek	Retrieve a Wikipedia paragraph that provides an answer to the given query about the image. Determine the Wikipedia snippet that matches the question of this image. I want to find a paragraph from Wikipedia that answers my question about this image. You have to find a Wikipedia segment that answers the question about the displayed image.
$(q^i, q^t) \rightarrow c^i$	FashionIQ	Find a fashion image that aligns with the reference image and style note. With the reference image and modification instructions, find the described fashion look. Given the reference image and design hint, identify the matching fashion image. I'm looking for a similar fashion product image with the described style changes.
	CIRR	Retrieve a day-to-day image that aligns with the modification instructions of the provided image. Pull up a common scene image like this one, but with the modifications I asked for. Can you help me find a daily image that meets the modification from the given image? I'm looking for a similar everyday image with the described changes.
$(q^i, q^t) \rightarrow (c^i, c^t)$	OVEN	Retrieve a Wikipedia image-description pair that provides evidence for the question of this image. Determine the Wikipedia image-snippet pair that clarifies the entity in this picture. I want to find an image and subject description from Wikipedia that answers my question about this image. I want to know the subject in the photo. Can you provide the relevant Wikipedia section and image?
	InfoSeek	Retrieve a Wikipedia image-description pair that provides evidence for the question of this image. Determine the Wikipedia image-snippet pair that matches my question about this image. I want to find an image and subject description from Wikipedia that answers my question about this image. I want to address the query about this picture. Please pull up a relevant Wikipedia section and image.

Table 4. Summary of the M-BEIR instructions.

E. Additional Experimental Results

E.1. Experimental Results on the M-BEIR in Global Pool Setting

The M-BEIR benchmark can be evaluated in two distinct settings: the global pool and the local pool. The key difference between these settings lies in the composition of the candidate pool, which is either constructed from all datasets collectively or restricted to the specific dataset currently under evaluation. In the main paper, we report results using the **local pool setting**. This section provides supplementary evaluation results based on the **global pool setting**.

The experimental results are presented in Table 5. Our method also demonstrates exceptional performance under the global pool setting, achieving an average of 12.5 points higher than UniIR-CLIP.

Methods	$q^t \rightarrow c^i$			$q^t \rightarrow c^t$		$q^t \rightarrow (c^i, c^t)$		$q^i \rightarrow c^t$			$q^i \rightarrow c^i$		$(q^i, q^t) \rightarrow c^t$		$(q^i, q^t) \rightarrow c^i$		$(q^i, q^t) \rightarrow (c^i, c^t)$		Avg.
	VN	COCO	F200K	WebQA	EDIS	WebQA	VN	COCO	F200K	NIGHTS	OVEN	InfoS	FIQ	CIRR	OVEN	InfoS			
	R@5	R@5	R@10	R@5	R@5	R@5	R@5	R@5	R@10	R@5	R@5	R@5	R@10	R@5	R@5	R@5			
Supervised - Dual Encoder																			
UniIR-BLIP _{FF} [11]	23.0	75.6	25.4	79.5	50.3	79.7	21.1	88.8	27.6	33.0	38.7	19.7	28.5	51.4	57.8	27.7	45.5		
UniIR-CLIP _{SF} [11]	42.6	77.9	17.8	84.7	59.4	78.8	42.8	92.3	17.9	32.0	39.2	24.0	24.3	43.9	60.2	44.6	48.9		
Supervised - LMMs																			
LamRA-Ret	41.3	75.4	28.7	85.8	62.5	81.0	39.3	90.4	30.4	32.1	48.4	48.7	33.1	50.5	70.0	60.0	54.9		
LamRA	46.9	78.0	32.5	96.5	74.4	87.1	47.6	92.4	36.6	34.2	54.0	58.7	37.4	59.7	72.6	74.0	61.4		

Table 5. **Comparison with up-to-date state-of-the-arts on M-BEIR test set in global pool setting.** The first row indicates the retrieval task type: q^t for text queries, q^i for image queries, c^t for text candidates, and c^i for image candidates. Abbreviations used include VN for VisualNews, F200K for Fashion200K, InfoS for InfoSeek, and FIQ for FashionIQ. Evaluation standards follow UniIR, with FashionIQ and Fashion200K using Recall@10, while all other evaluations employ Recall@5.

E.2. Qwen2-VL Vs. Qwen2.5-VL

In Table 6 and Table 7, we present a comparative analysis of the performance between Qwen2-VL and Qwen2.5-VL. It is evident that our framework demonstrates commendable performance across both models, thereby underscoring the versatility and generalizability of our proposed framework.

E.3. Detailed Experimental Results on the Pointwise and Listwise Reranking

As shown in Table 8, we present a comprehensive comparison of LamRA-Rank’s pointwise and listwise reranking methods across a range of tasks. Both approaches demonstrate enhanced performance in various applications. However, with respect to inference time, the listwise reranking method does not consistently outperform the pointwise approach, particularly when the candidate set includes images. This discrepancy arises because current LMMs often require hundreds of tokens to represent a single image, significantly increasing the context length during listwise reranking. The extended context can adversely affect inference speed. Ongoing research [1, 5] is actively investigating methods to reduce the number of visual tokens required. We anticipate that, as LMMs continue evolving, the use of LMMs for listwise reranking will become an increasingly prevalent approach.

Methods	$q^t \rightarrow c^i$			$q^t \rightarrow c^t$		$q^t \rightarrow (c^i, c^t)$		$q^i \rightarrow c^t$			$q^i \rightarrow c^i$		$(q^i, q^t) \rightarrow c^t$		$(q^i, q^t) \rightarrow c^i$		$(q^i, q^t) \rightarrow (c^i, c^t)$		Avg.
	VN	COCO	F200K	WebQA	EDIS	WebQA	VN	COCO	F200K	NIGHTS	OVEN	InfoS	FIQ	CIRR	OVEN	InfoS			
	R@5	R@5	R@10	R@5	R@5	R@5	R@5	R@5	R@10	R@5	R@5	R@5	R@10	R@5	R@5	R@5			
Qwen2-VL-7B																			
LamRA-Ret	41.6	81.5	28.7	86.0	62.6	81.2	39.6	90.6	30.4	32.1	54.1	52.1	33.2	53.1	76.2	63.3	56.6		
LamRA	48.0	85.2	32.9	96.7	75.8	87.7	48.6	92.3	36.1	33.5	59.2	64.1	37.8	63.3	79.2	78.3	63.7		
Qwen2.5-VL-7B																			
LamRA-Ret	38.5	81.1	24.0	85.8	59.5	81.1	36.3	91.2	23.2	30.9	58.4	57.3	32.0	52.3	79.7	65.1	56.0		
LamRA	45.9	85.2	27.3	96.2	73.0	87.8	45.4	92.8	27.1	33.6	61.6	68.6	36.2	63.5	80.9	79.9	62.8		

Table 6. Performance comparison between Qwen2-VL and Qwen2.5-VL on the M-BEIR test set.

Methods	$q^t \rightarrow c^i$			$q^i \rightarrow c^t$			$(q^i, q^t) \rightarrow c^i$		$q^{\text{dialog}} \rightarrow c^i$	$(q^i \oplus q^t) \rightarrow c^i$		ITM	
	Share4V	Urban*	Flickr	Share4V	Urban*	Flickr	CIRCO*	GeneCIS*	VisD*	VIST	MT-FIQ*	CC-Neg	Sugar-Crepe*
	R@1	R@1	R@1	R@1	R@1	R@1	MAP@5	R@1	R@1	R@1	R@5	Acc.	Acc.
<i>Qwen2-VL-7B</i>													
LamRA-Ret	93.3	95.1	82.8	88.1	94.3	92.7	33.2	18.9	62.8	23.1	60.9	79.6	85.8
LamRA	97.9	98.8	88.1	96.5	98.0	97.6	42.8	24.8	70.9	28.6	63.9	85.9	93.5
<i>Qwen2.5-VL-7B</i>													
LamRA-Ret	93.8	95.6	82.7	92.9	96.0	93.3	34.4	19.2	64.1	22.8	60.7	78.0	86.4
LamRA	97.5	99.1	88.6	97.4	98.3	96.7	43.0	23.6	73.5	30.0	62.2	79.5	92.8

Table 7. Performance comparison between Qwen2-VL and Qwen2.5-VL on unseen dataset.

Task	LamRA-Ret	LamRA-Rank(P)		LamRA-Rank(L)	
	R@1	R@1	Time	R@1	Time
$q^t \rightarrow c^i$	29.7	33.2	0.041s	33.1	0.057s
$q^t \rightarrow c^t$	58.2	75.9	0.020s	75.9	0.010s
$q^t \rightarrow (c^i, c^t)$	41.7	50.9	0.044s	50.5	0.099s
$q^i \rightarrow c^t$	34.0	38.5	0.043s	37.9	0.012s
$(q^i, q^t) \rightarrow c^i$	18.5	24.5	0.071s	24.3	0.067s
$q^i \rightarrow c^i$	8.4	10.0	0.046s	8.5	0.048s
$(q^i, q^t) \rightarrow c^t$	30.1	37.3	0.047s	36.6	0.017s
$(q^i, q^t) \rightarrow (c^i, c^t)$	33.4	39.9	0.084s	39.5	0.085s

Table 8. Detailed Comparison of Recall@1 performance and inference costs between pointwise (LamRA-Rank(P)) and listwise (LamRA-Rank(L)) reranking methods on M-BEIR. Reranking is applied to the top-5 results. Time denotes the per-query inference time cost measured on eight A100 GPUs with a batch size of 32.

F. Exploration of RAG Applications

According to the analysis in Section 4.2 of the main paper, the LMM demonstrates significant potential to serve as a universal retriever. This raises an important question: *Can the retrieval and generative capabilities be integrated within the same LMM?*

We have conducted experiments on three Knowledge-based Visual Question Answering (KVQA) tasks. KVQA requires to first retrieve the relevant documents, then answer the associated questions with the retrieved information. Specifically, we train the retrieval and VQA tasks simultaneously using LoRA during the training process. The experimental results, presented in Table 9, indicate that the retrieval performance of our method on the three datasets surpasses the current SOTA. Furthermore, the accuracy of answering questions based on the retrieved documents is comparable to or exceeds the current SOTA. These findings demonstrate the feasibility of integrating retrieval and generative capabilities within a single LMM, which we consider a promising direction for future work.

Method	OKVQA [8]	Infoseek [3]	E-VQA [9]
<i>Retrieval (PR@5)</i>			
PreFLMR [7]	70.9	62.1	73.7
Ours	89.0	73.4	75.0
<i>VQA (ACC.)</i>			
RA-VQAv2 w/ PreFLMR [6]	61.9	30.7	54.5
Ours	64.3	28.8	56.2

Table 9. Comparison with state-of-the-art (SOTA) Methods on KVQA tasks.

G. Limitations & Future Work

While our method has demonstrated superior performance over existing models, there remain several limitations that can be further improved. (i) The current implementation of LamRA necessitates training separate sets of LoRA parameters for the retrieval and reranking tasks. In the future, we may explore strategies for jointly training these two tasks simultaneously or consider integrating retrieval training into the SFT stage. (ii) Due to constraints in context length and computation resources, the listwise reranking method of LamRA-Rank currently supports input from 2-5 candidates. Investigating strategies to enable support for a larger number of candidates is an important area for future research. (iii) The current model sizes are restricted to 2B and 7B parameters. Further exploration of LMMs for multimodal information retrieval holds significant potential for improvement. (iv) The current model may exhibit sensitivity to instructions, where variations in instructions can lead to certain fluctuations in performance. Training a model that is more robust to variations in instructions is a promising direction for future research.

H. More Qualitative Results

H.1. Successful Cases

In this section, we show additional examples of successful retrieval. As illustrated in Figure 1 through 9, our method effectively handles a diverse range of retrieval tasks.

H.2. Failure Cases

We present several failure cases in Figure 10. As observed, some failures are attributable to false-negative candidates, while others stem from inherently challenging queries that lead to retrieval failures, as exemplified in the last row.

Query Cues	Retrieved Results		
<p>Instruction: I want you to retrieve an image of this news caption.</p> <p>An officer checks a trash can near a transit stop a from police headquarters.</p> $q^t \rightarrow c^i$			
<p>Instruction: Find me an everyday image that matches the given caption.</p> <p>A motorcycle sits parked across from a herd of livestock.</p> $q^t \rightarrow c^i$			
<p>Instruction: Match the provided description to the correct fashion item photo.</p> <p>Brown studded zipped jacket.</p> $q^t \rightarrow c^i$			

Figure 1. Qualitative examples on text-to-image retrieval task, where the red box marks the ground truth.

Query Cues	Retrieved Results		
<p>Instruction: Retrieve passages from Wikipedia that provide answer to the question.</p> <p>What species of fruit bat is found in locations that are further south ...</p> $q^t \rightarrow c^t$	Veldkamp's dwarf epauletted fruit bat is a species of bat in the ...	Wahlberg's epauletted fruit bat is a species of megabat ...	Roost trees may be shared with other species, though roosting ...
<p>Instruction: You have to find a Wikipedia that provide answer to the question.</p> <p>Which bird is more likely to survive in a higher altitude ...</p> $q^t \rightarrow c^t$	The rufous-sided broadbill is a species of bird in the family ...	The rufous-sided broadbill ... Its natural habitat is subtropical ...	The rufous-sided warbling finch is a species of bird in the family ...
<p>Instruction: I am looking for a Wikipedia snippet that answer this question.</p> <p>Which occurred first, the last ever Daily Show hosted by Jon Stewart, or the retirement ...</p> $q^t \rightarrow c^t$	Joseph S. Benigno is an American sports radio personality. He was a ...	On February 10, 2015, Stewart announced that he would be ...	When Jon Stewart became the program's host in 1999 ...

Figure 2. Qualitative examples on text-to-text retrieval task, where the red text marks the ground truth.










Query Cues	Retrieved Results		
<p>Instruction: Find a news image that matches the provided caption.</p> <p>Tom Holland makes his debut in the Spidey suit in Captain America Civil War.</p> <p>$q^t \rightarrow (c^i, c^t)$</p>	 <p>Tom Holland ...</p>	 <p>... for spider ...</p>	 <p>... new suit ...</p>
<p>Instruction: Identify the news photo for the given caption.</p> <p>Antigovernment protesters gather in front of the Democracy Monument during a rally in Bangkok Thailand Friday.</p> <p>$q^t \rightarrow (c^i, c^t)$</p>	 <p>... Bangkok rally ...</p>	 <p>... never surrender ...</p>	 <p>Thai capital hit by ...</p>
<p>Instruction: Find a Wikipedia image that answers this question.</p> <p>What shapes are painted down the center of Salou Boulevard?</p> <p>$q^t \rightarrow (c^i, c^t)$</p>	 <p>Salou boulevard.</p>	 <p>43840, Salou, ...</p>	 <p>Barcelona street ...</p>

Figure 3. Qualitative examples on text-to-text-image retrieval task, with the ground truth indicated by a red box and red text.




Query Cues	Retrieved Results		
 <p>Instruction: Find a caption for the news in the given photo.</p> <p>$q^i \rightarrow c^t$</p>	<p>Residents of a village ... with smartphones and a laptop.</p>	<p>A young woman enjoys social networking by using her ...</p>	<p>High volume low costs for India's cell users.</p>
 <p>Instruction: Retrieve the caption for the day-to-day image.</p> <p>$q^i \rightarrow c^t$</p>	<p>A woolly sheep stands in the grass looking at the camera.</p>	<p>A very large sheep is standing in the grass.</p>	<p>A shaggy haired sheep looking up in the field.</p>
 <p>Instruction: Find a product description for the fashion item.</p> <p>$q^i \rightarrow c^t$</p>	<p>Black long sleeve dress lace insert.</p>	<p>Black sheer panel dress.</p>	<p>Black sheer long sleeve midi dress.</p>

Figure 4. Qualitative examples on image-to-text retrieval task, where the red text marks the ground truth.

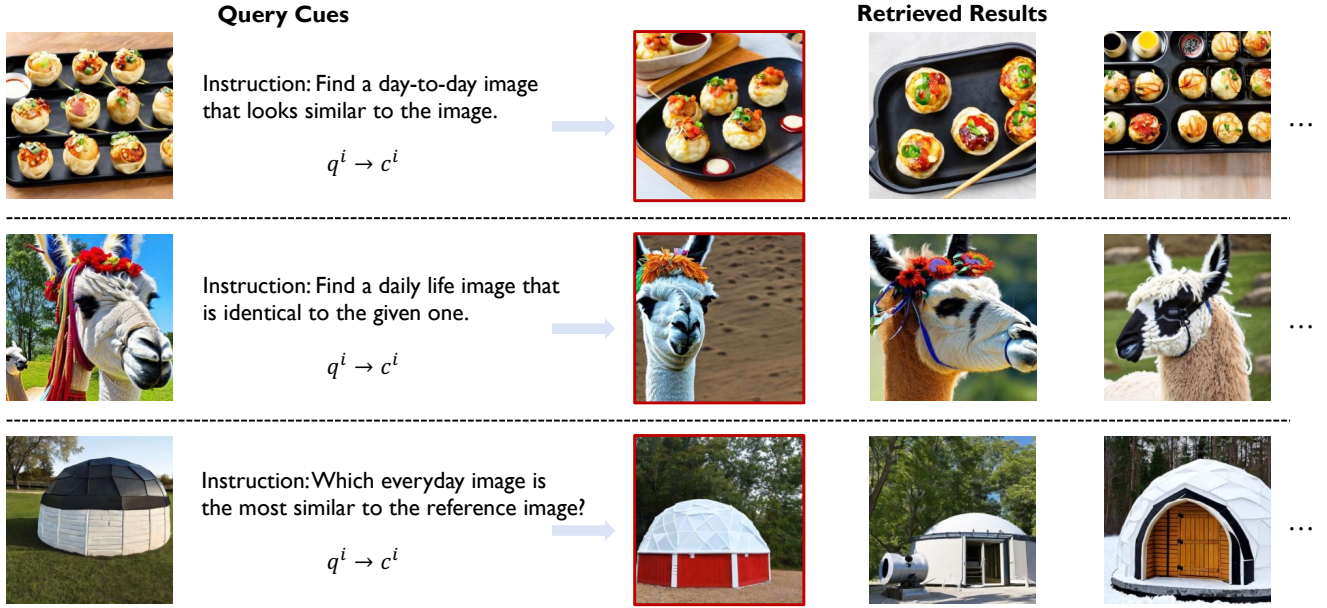


Figure 5. Qualitative examples on image-to-image retrieval task, where the red box marks the ground truth.

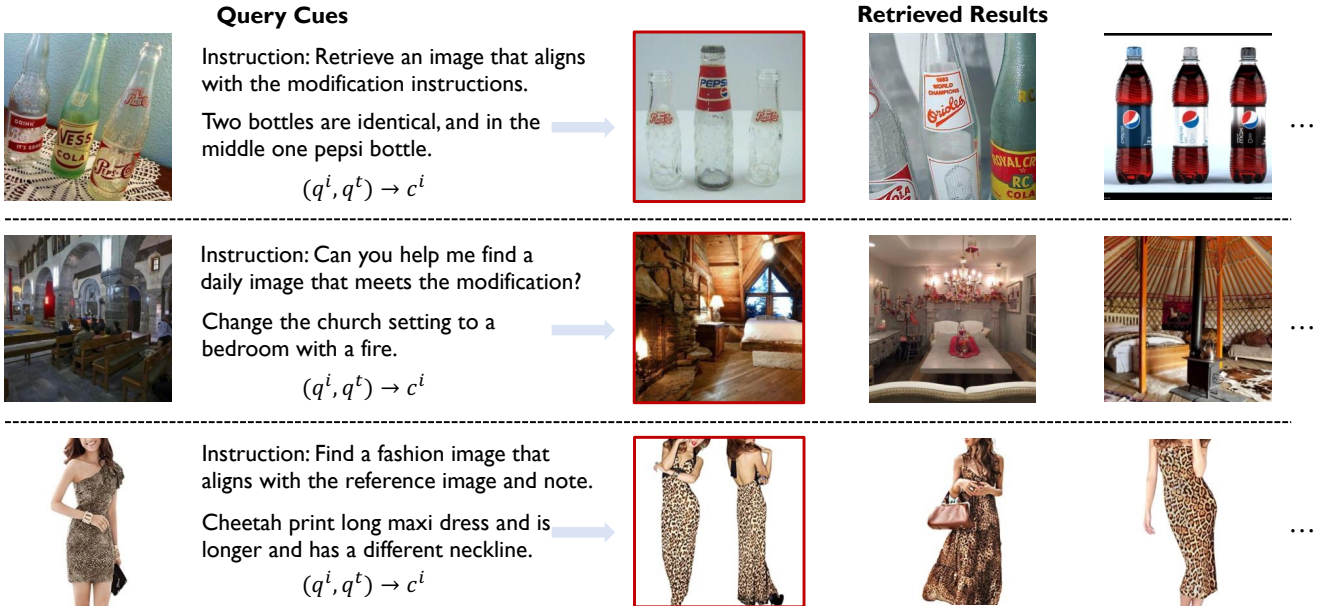


Figure 6. Qualitative examples on text-image-to-image retrieval task, where the red box marks the ground truth.




Query Cues		Retrieved Results	
	<p>Instruction: Retrieve a Wikipedia snippet that provides an answer to the query.</p> <p>What kind of plant is this?</p> <p>$(q^i, q^t) \rightarrow c^t$</p>	<p>Festuca. Festuca is a genus of flowering plants belonging to ...</p>	<p>Ornamental grass are grown as ornamental plants. Ornamental ...</p> <p>Festuca pallens, the blue fescue, is a species of grass. ...</p>
	<p>Instruction: Determine the Wikipedia snippet that identifies the visual entity.</p> <p>What kind of food is this?</p> <p>$(q^i, q^t) \rightarrow c^t$</p>	<p>Pumpkin seed. A pumpkin seed, also known in North America as ...</p>	<p>Sunflower seed. The sunflower seed is the seed of the sunflower ...</p> <p>List of culinary nuts. A culinary nut is a dry, edible fruit or seed ...</p>
	<p>Instruction: You have to find a Wikipedia segment that answers the question.</p> <p>Which city or region does this building locate in?</p> <p>$(q^i, q^t) \rightarrow c^t$</p>	<p>Guarda Cathedral. In Guarda Cathedral was finished around 1540. Around ...</p>	<p>Viseu Cathedral. Santos Pacheco, who was also responsible for ...</p> <p>Castle of Braganca. The ports of the Princess quarters, pretending to ...</p>

Figure 7. Qualitative examples on text-image-to-text retrieval task, where the red text marks the ground truth.












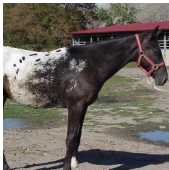
Query Cues		Retrieved Results		
	<p>Instruction: Retrieve a Wikipedia image-description pair that provides evidence ...</p> <p>Which category of food is shown in the image?</p> <p>$(q^i, q^t) \rightarrow (c^i, c^t)$</p>			
		Pina colada is a cocktail ...	Eggnog ...	Bombardino ...
	<p>Instruction: Determine the Wikipedia image-snippet pair that clarifies ...</p> <p>What is this building called?</p> <p>$(q^i, q^t) \rightarrow (c^i, c^t)$</p>			
		Stonehenge ...	Stones of Stenness...	Killin Stone Circle ...
	<p>Instruction: I want to find an image and subject description that answers ...</p> <p>What is shown in the photo?</p> <p>$(q^i, q^t) \rightarrow (c^i, c^t)$</p>			
		Mustang ...	Marismeño ...	Nez Perce Horse ...

Figure 8. Qualitative examples on text-image-to-text-image retrieval task, with the ground truth indicated by a red box and red text.

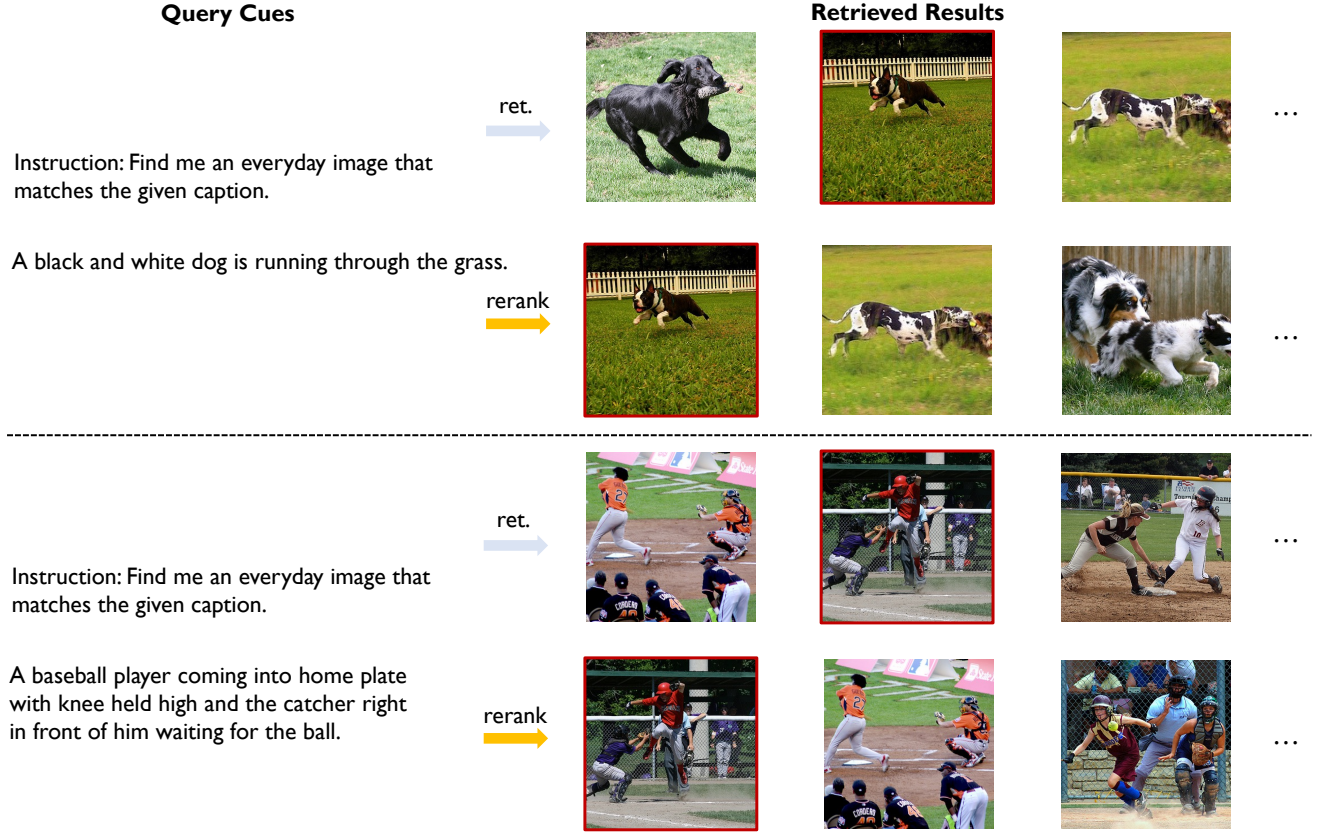


Figure 9. Some examples of retrieval followed by reranking are provided, with the red box indicating the ground truth.

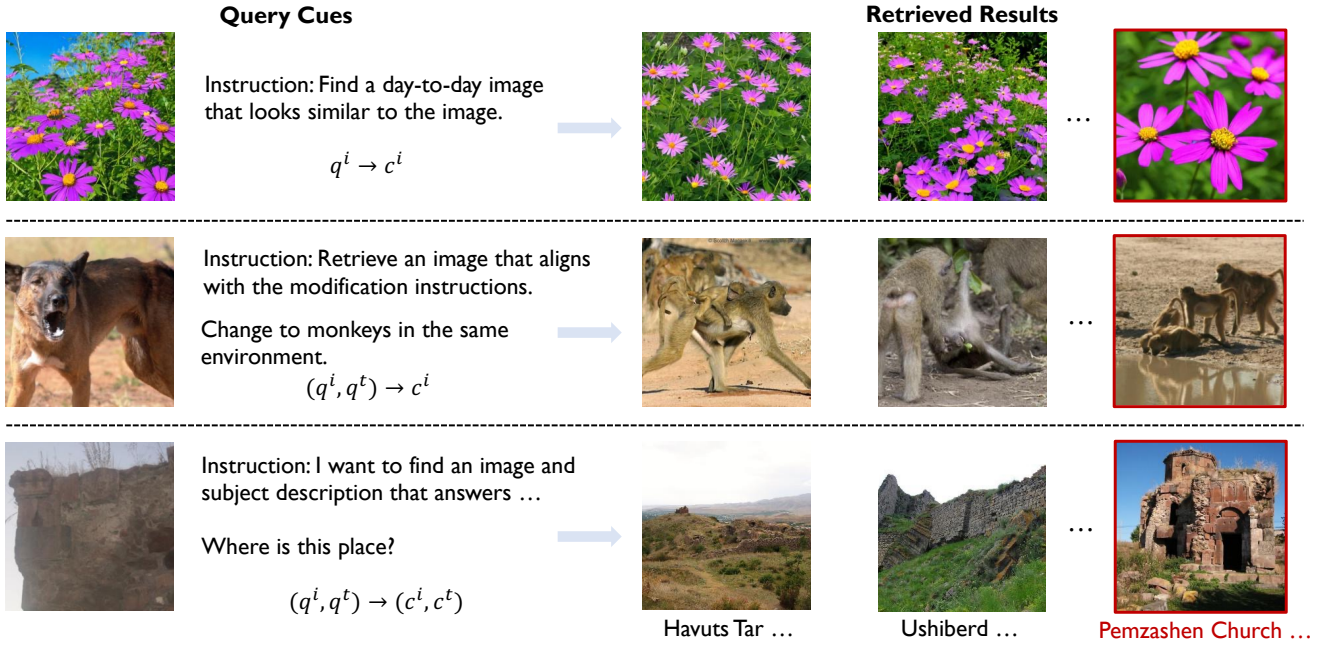


Figure 10. Some failure cases.

References

- [1] Jieneng Chen, Luoxin Ye, Ju He, Zhao-Yang Wang, Daniel Khashabi, and Alan Yuille. Efficient large multi-modal models via visual context compression. In *Advances in Neural Information Processing Systems*, 2024. 4
- [2] Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. Sharegpt4v: Improving large multi-modal models with better captions. In *Proceedings of the European Conference on Computer Vision*, 2024. 1
- [3] 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 Conference on Empirical Methods in Natural Language Processing*, 2023. 6
- [4] Tianyu Gao, Xingcheng Yao, and Danqi Chen. Simcse: Simple contrastive learning of sentence embeddings. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2021. 1
- [5] Wentong Li, Yuqian Yuan, Jian Liu, Dongqi Tang, Song Wang, Jie Qin, Jianke Zhu, and Lei Zhang. Tokenpacker: Efficient visual projector for multimodal llm. *arXiv preprint arXiv:2407.02392*, 2024. 4
- [6] Weizhe Lin, Jinghong Chen, Jingbiao Mei, Alexandru Coca, and Bill Byrne. Fine-grained late-interaction multi-modal retrieval for retrieval augmented visual question answering. In *Advances in Neural Information Processing Systems*, 2023. 6
- [7] Weizhe Lin, Jingbiao Mei, Jinghong Chen, and Bill Byrne. PreFLMR: Scaling up fine-grained late-interaction multi-modal retrievers. In *Association for Computational Linguistics*, 2024. 6
- [8] Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2019. 6
- [9] Thomas Mensink, Jasper Uijlings, Lluís Castrejon, Arushi Goel, Felipe Cadar, Howard Zhou, Fei Sha, André Araujo, and Vittorio Ferrari. Encyclopedic vqa: Visual questions about detailed properties of fine-grained categories. In *Proceedings of the International Conference on Computer Vision*, 2023. 6
- [10] OpenAI. Gpt-4v(ision) system card, 2023. 2
- [11] Cong Wei, Yang Chen, Haonan Chen, Hexiang Hu, Ge Zhang, Jie Fu, Alan Ritter, and Wenhui Chen. Uniir: Training and benchmarking universal multimodal information retrievers. In *Proceedings of the European Conference on Computer Vision*, 2024. 2, 4
- [12] Kai Zhang, Yi Luan, Hexiang Hu, Kenton Lee, Siyuan Qiao, Wenhui Chen, Yu Su, and Ming-Wei Chang. Magiclens: Self-supervised image retrieval with open-ended instructions. In *Proceedings of the International Conference on Machine Learning*, 2024. 1
- [13] Haozhe Zhao, Xiaojian Ma, Liang Chen, Shuzheng Si, Rujie Wu, Kaikai An, Peiyu Yu, Minjia Zhang, Qing Li, and Baobao Chang. Ultraedit: Instruction-based fine-grained image editing at scale. In *Advances in Neural Information Processing Systems*, 2024. 1