MegaPairs: Massive Data Synthesis for Universal Multimodal Retrieval

Junjie Zhou[1,2*], **Yongping Xiong**[1*], **Zheng Liu**[2,5†], **Ze Liu**[3,2], **Shitao Xiao**[2],
Yueze Wang[2], **Bo Zhao**[4], **Chen Jason Zhang**[5], **Defu Lian**[3†]

[1] State Key Laboratory of Networking and Switching Technology,
Beijing University of Posts and Telecommunications
[2] Beijing Academy of Artificial Intelligence
[3] University of Science and Technology of China
[4] Shanghai Jiao Tong University, [5] The Hong Kong Polytechnic University
{junjiebupt, zhengliu1026}@gmail.com, liandefu@ustc.edu.cn

Abstract

Despite the rapidly growing demand for multimodal retrieval, progress in this field remains severely constrained by a lack of training data. In this paper, we introduce **MegaPairs**, a novel data synthesis method that leverages vision language models (VLMs) and open-domain images, together with a massive synthetic dataset generated from this method. Our empirical analysis shows that MegaPairs generates high-quality data, enabling the multimodal retriever to significantly outperform the baseline model trained on $70\times$ more data from existing datasets. Moreover, since MegaPairs solely relies on general image corpora and open-source VLMs, it can be easily scaled up, enabling continuous improvements in retrieval performance. In this stage, we produced more than 26 million training instances and trained several models of varying sizes using this data. These new models achieve state-of-the-art zero-shot performance across 4 popular composed image retrieval (**CIR**) benchmarks and the highest overall performance on the 36 datasets provided by **MMEB**. They also demonstrate notable performance improvements with additional downstream fine-tuning. Our code, synthesized dataset, and pre-trained models are publicly available at https://github.com/VectorSpaceLab/MegaPairs.

1 Introduction

Multimodal retrieval is a critical research problem for IR and AI communities. It aims to satisfy people's information needs across different data modalities, especially texts and images. Nowadays, multimodal retrieval has been applied to a wide variety of real-world scenarios, such as image search (Chen et al., 2015; Wu et al., 2021; Zhang et al., 2024), visual question answering (VQA) (Marino et al., 2019; Mathew et al., 2021),

and retrieval-augmented generation (RAG) of vision language models (Chen et al., 2022; Yu et al., 2024). Given the widespread application scenarios, it's necessary to develop universal multimodal retrievers which can uniformly support any task requirements and working domains.

The progress of universal multimodal retrievers have been substantially advanced on top of the pre-trained vision-languages models, like CLIP (Radford et al., 2021), ALIGN (Jia et al., 2021), and SigLIP (Zhai et al., 2023). These models are pre-trained to produce discriminative and unified representations for texts and images, thus creating a solid foundation for multimodal retrieval. However, the existing vision-language encoders are mostly pre-trained from text-image matching tasks. Although these models have achieved an initial capability for text-to-image retrieval (Young et al., 2014; Chen et al., 2015), they are insufficient for other common multimodal tasks, such as composed image retrieval (Liu et al., 2021; Baldrati et al., 2023; Zhang et al., 2024) and multimodal document retrieval (Chang et al., 2022; Liu et al., 2022).

To enhance the multi-task capacity, fine-tuning pre-trained models with comprehensive instructions, commonly known as instruction-tuning, has gained significant popularity. This approach was first applied in the supervised fine-tuning of large language models (LLMs) (Ouyang et al., 2022; Wei et al., 2021; Chung et al., 2024), and later introduced for training text embeddings (Su et al., 2022; Asai et al., 2022; Zhang et al., 2023; Xiao et al., 2024). Building on these successes, instruction-tuning has been further extended to multimodal embedding models (Wei et al., 2024; Sharifymoghaddam et al., 2024), where pre-trained vision-language encoders are continually fine-tuned using a variety of multimodal retrieval instructions. Given the scarcity of instruction-tuning data for embedding models, researchers have proposed leveraging LLMs to generate synthetic data from In-

---

*Co-first authors.
†Corresponding authors.

19076

ternet resources (Wang et al., 2023). In the field of multimodal retrieval, a notable example is presented by MagicLens (Zhang et al., 2024), which synthesizes open-ended search instructions for co-existing images within the same webpage.

Despite recent advancements by MagicLens, current data synthesis methods still face significant limitations in data *scalability*, *quality*, *diversity*, and *availability*. Specifically, only a small fraction of webpages on the internet contain multiple images (scalability), not to mention that many of these co-existing images are either unrelated or near-duplicates (quality). Besides, the remaining correlated images often exhibit monotonous relationships, such as different angles of the same object (diversity). Finally, large-scale instruction-tuning datasets for multimodal retrieval are typically held privately by individual research labs (availability).

In this paper, we introduce a novel data synthesis method called **MegaPairs**, accompanied by a large-scale instruction dataset generated using this approach. MegaPairs is distinguished by its construction of *a heterogeneous KNN triplet for open-domain images*. Particularly, it leverages three different similarity models to sample correlated image pairs, including CLIP vision-encoder for visual-semantic correlations (Sun et al., 2023), DINO vision-encoder for visual-pattern correlations (Oquab et al., 2024), and CLIP text-encoder for caption correlations. The sampled image pairs are presented for the VLM and LLM annotators, which generate comprehensive descriptions of the relationships between the two images and create pseudo-retrieval instructions based on the descriptions. This approach enables a huge amount of instructions to be generated for a general dataset, like Datacomp (Gadre et al., 2024), which significantly improves the scalability of data synthesis. It also introduces diverse instructions of guaranteed quality, given its sampling of heterogeneous relationships from open-ended image corpora. Additionally, by utilizing open-source VLM and LLM models (e.g., InternVL2-26B (Chen et al., 2024b), Llama-3-8B (Dubey et al., 2024)), the entire process can operate at a low cost.

We've produced 26 million data instances in this stage, achieving superior data quality compared to the existing datasets. In our pilot experiment, with just 500K sampled instances from MegaPairs, the same pre-trained model's fine-tuning performance already surpasses that of the entire 36.7M training instances from MagicLens, i.e., delivering better re-sults with $70\times$ less training data. We further trained three multimodal retrievers, **MMRet**, of varying sizes based on the whole synthetic dataset and perform comprehensive evaluations with a wide range of multimodal retrieval tasks. Remarkably, MMRet achieved state-of-the-art performance on 4 popular composed image retrieval (CIR) benchmarks and the 36 datasets provided by MMEB (Jiang et al., 2024b) in the zero-shot setting. Furthermore, the models demonstrated substantial improvements and maintain leading positions after downstream fine-tuning. The entire suite of assets, including the dataset, the well-trained models, and the data production pipeline, will be made publicly available to advance the future progress in this field.

2 Related Work

Multimodal Retrieval. Traditionally, retrieval tasks have focused on scenarios where queries and candidates exist in distinct modalities, such as unimodal retrieval (Thakur et al., 2021) and cross-modal retrieval (Chen et al., 2015). However, there is a growing demand for multimodal retrieval tasks, where queries or candidates integrate both image and text modalities. These tasks have wide applications, including image retrieval with instructions (Wu et al., 2021; Liu et al., 2021; Zhang et al., 2024), multimodal document retrieval (Chang et al., 2022; Liu et al., 2022), knowledge retrieval with multimodal queries (Luo et al., 2023), and retrieval-augmented generation (Yasunaga et al., 2023; Yu et al., 2024). Most existing methods employ pre-trained vision-language models (VLMs) to address these tasks (Radford et al., 2021; Li et al., 2023; Saito et al., 2023). However, the common VLMs are purely trained on image-text matching datasets (Changpinyo et al., 2021; Schuhmann et al., 2022), which are in lack of ability to jointly encode and comprehend both modalities effectively. As a result, it is necessary to create proper datasets so as to extend VLMs for the diversified multimodal retrieval tasks.

Instruction Tuning for Multimodal Retrieval. Instruction-tuning is a popular strategy to enhance the multi-task capacity for both large language models (Ouyang et al., 2022; Chung et al., 2024; Tan et al., 2024) and embedding models (Su et al., 2022; Asai et al., 2022; Zhang et al., 2023; Xiao et al., 2024; Chen et al., 2024a). While there have been a few instruction datasets proposed for mul-timodal retrieval (Liu et al., 2021, 2022; Chang
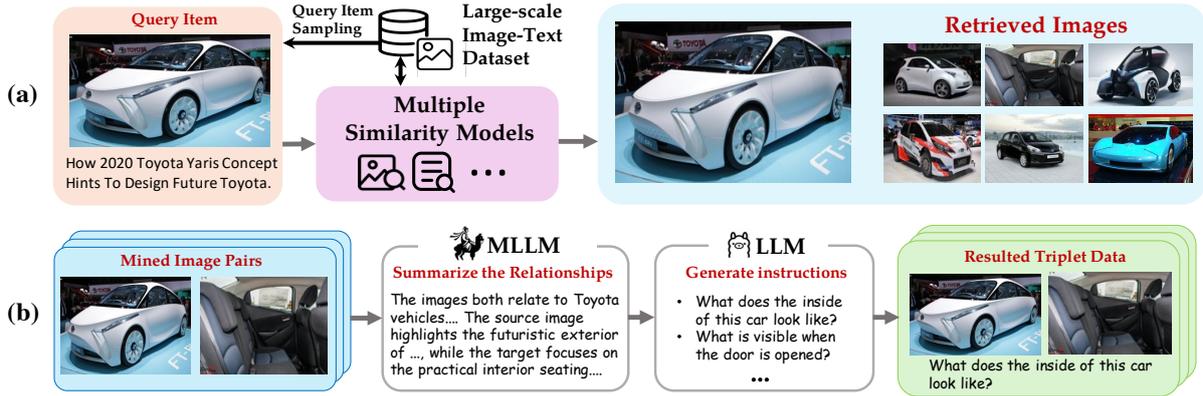
Figure 1: Construction pipeline of multimodal triplets: **(a)** mining of image pairs, **(b)** generation of open-ended instructions. Multiple similarity models are used to introduce diversified correlations for the image pairs.

et al., 2022; Wei et al., 2024; Zhou et al., 2024), they are limited in scale and diversity due to their reliance on human annotation. Recently, a notable progress was made by MagicLens (Zhang et al., 2024), where a large-scale open-ended search instruction dataset is created from the co-existed images within webpages. However, given the shortage of multi-image webs, MagicLens is limited by its scalability and data-quality. Moreover, this dataset is still held private and inaccessible to public users. As a result, the creation and release of high-quality instruction-tuning datasets have become imperative for advancing multimodal retrieval research.

3 Methodology

3.1 MegaPairs Construction

Training on large-scale open-world data significantly enhances the generalization capabilities of foundation models. For instance, CLIP (Radford et al., 2021) has achieved remarkable advancements in cross-modal retrieval and various downstream tasks due to its extensive training on text-image pairs. However, the multimodal instruction tuning data, despite its importance to multimodal retrieval (Zhang et al., 2024), is scarce in natural world and expensive to annotate by human effort. In this paper, we propose to construct large-scale multimodal instruction-tuning datasets through data synthesis. Formally, each data instance contains the following triplet: a pair of images $(\mathcal{I}_q, \mathcal{I}_t)$, together with a textual instructions $\mathcal{T}_{q \to t}$ specifying the transition relationship from query image $\mathcal{I}_q$ to target image $\mathcal{I}_t$.

We identify two primary technical challenges in acquiring such triplets: (1) *sampling relevant and diversified image pairs at scale*, (2) *precise annotation of instruction for the sampled image*

pair. To address these challenges, we propose leveraging the common open-domain image corpora. Intuitively, a large-scale corpus contains abundant correlated images of diverse semantic relationships, which can be mined and annotated for our instruction-tuning data. Our data synthesis pipeline is demonstrated as Figure 1, which involves two main components: the mining of image pairs and the generation of open-ended instructions.

Mining Correlated Image Pairs. As illustrated in Figure 1(a), we propose sampling correlated image pairs from a large-scale image corpus. For each query image $(\mathcal{I}_q, \mathcal{C}_q)$, we utilize multiple similarity models to search for a diverse set of correlated target images of heterogeneous correlations $\{\mathcal{I}_{t_1}, \mathcal{I}_{t_2}, \ldots, \mathcal{I}_{t_n}\}$. In our work, the following types of correlations are used: (1) visual-semantic correlation, which measures the semantic correlation of two images regardless of visual similarity, e.g., two different views of the same cars; (2) visual-pattern correlation, which captures the visual similarity of two images regardless of semantic correlation, e.g., different cars in similar backgrounds; (3) caption correlation, which measures the textual similarity between two images' captions.

Recognizing the importance of **hard negatives** in training retrieval models (Xiong et al., 2020; Hofstätter et al., 2021; Zhang et al., 2022), for each pair $(\mathcal{I}_q, \mathcal{I}_{t_i})$, we include additional images $\{\mathcal{I}_{t_j} \mid t_j \neq t_i\}$ from the retrieved set as hard negative samples. This approach is simple but empirically effective. We validate the scalability and quality of our data pairs in Section 4.3, with additional examples visualized in Appendix F.

Generating Open-Ended Instructions. As shown in Figure 1(b), we utilize open-source multimodal large language models (MLLM) and

large language models (LLM) for the automated annotation of mined image pairs $\mathcal{P} = \{(\mathcal{I}_q, \mathcal{I}_{t_i})\}$. Initially, each image pair $(\mathcal{I}_q, \mathcal{I}_{t_i})$ is processed by the MLLM to generate a detailed description $\mathcal{D}_i$ of the common concepts and differences between the query image $\mathcal{I}_q$ and the target image $\mathcal{I}_{t_i}$. This description $\mathcal{D}_i$ is then refined by the LLM to produce textual instructions $\mathcal{T}_{q \to t_i}$. We prompt the LLM to generate multiple $\mathcal{T}_{q \to t_i}$ for each pair, enhancing the diversity of the textual instructions. Ultimately, we construct a multimodal triplet $(\mathcal{I}_q, \mathcal{T}_{q \to t_i}, \mathcal{I}_{t_i})$, where $(\mathcal{I}_q, \mathcal{T}_{q \to t_i})$ can be used to retrieve $\mathcal{I}_{t_i}$. This two-step annotation method ensures both accuracy and diversity in the automated annotation process while leveraging open-source models. The detailed prompts can be found in Appendix A.

Implementations. A dataset of 26,235,105 image pairs is created based on the above data synthesis pipeline. We utilize a subset from the Recap-DataComp-1B (Li et al., 2024b) as our image corpus, containing 20 million captioned images. For similarity models, we employ EVA-CLIP's image encoder for visual-semantic correlation (Sun et al., 2023), DINOv2 (Oquab et al., 2024) for visual-pattern correlation, and EVA-CLIP's text encoder for caption similarity. We filter the image pairs whose similarity score is within (0.8, 0.96), thus eliminating weak associations and near duplications. We further leverage InternVL2-26B (Chen et al., 2024b) and LLaMA3-8B (Dubey et al., 2024) to generate the open-ended instructions. For each image pair, we create at least three different textual instructions and introduce five hard negatives.

3.2 MMRet Model

We propose MMRet, a series of models designed for universal multimodal retrieval based on pre-trained vision-language models (VLMs). Our MM-Ret integrates two distinct VLM architectures to achieve a universal multimodal embedding.

CLIP-based MMRet. The original CLIP (Radford et al., 2021) model employs a dual encoder architecture that independently encodes image and text data. We denote the image encoder as Φ_I and the text encoder as Φ_T. Given an image I or text T, their embeddings are computed as follows:

$$\begin{aligned} \mathbf{e}_i &= \Phi_I(I) \\ \mathbf{e}_t &= \Phi_T(T) \end{aligned} \quad (1)$$

To produce the multimodal embedding for a composed image-text sample (I, T), we employ the

score-fusion strategy as used by UniIR (Wei et al., 2024), which directly uses an element-wise addition of the outputs from the dual encoders:

$$\mathbf{e}_{it} = \Phi_I(I) + \Phi_T(T) \quad (2)$$

In our CLIP-based MMRet, we trained both base and large models.

MLLM-based MMRet. The multimodal large language models (MLLMs) incorporate a visual encoder, typically based on a vision transformer (Dosovitskiy et al., 2021), into a large language model (LLM). This integration allows image tokens to be directly processed by the LLM. Consequently, MLLMs can effectively handle diverse multimodal inputs by converting any type of input into a sequence of tokens. For instance, composed image-text data is transformed into interleaved sequences of image and text tokens, enabling the model to process them seamlessly.

Our MMRet model builds upon the LLaVA-1.6 (Liu et al., 2024). In both training and inference stages, MMRet uses task-specific instructions for query inputs to improve generalization, aligning with standard practices in LLM-based embedding models (Wang et al., 2023; Li et al., 2024a). A typical multimodal query input is structured as follows:

$$\langle \text{instruct} \rangle \; \{\text{task\_inst}\} \; \langle \text{query} \rangle \; \{q_t\} \; \{q_i\} \; \texttt{[EOS]} \quad (3)$$

where {task_inst} represents the task-specific instruction, $\{q_t\}$ denotes the input query text, and $\{q_i\}$ is the input query image. The normalized last hidden state of the [EOS] token in the MLLM is used as the embedding of any given input sequence.

3.3 Multimodal Contrastive Learning

We employ multimodal contrastive learning to transform the original CLIP and MLLM into our MMRet model, enabling various multimodal retrieval tasks. We use the standard InfoNCE loss (Oord et al., 2018) as our training objective:

$$\mathcal{L} = -\frac{1}{|\mathcal{Q}|} \sum_{q_i \in \mathcal{Q}} \log \frac{\exp(\mathbf{e}_{q_i} \cdot \mathbf{e}_{c_i^+}/\tau)}{\sum_{c_j \in \mathcal{C}} \exp(\mathbf{e}_{q_i} \cdot \mathbf{e}_{c_j}/\tau)} \quad (4)$$

where the set $\mathcal{Q}$ includes all query samples q_i in a batch. The vectors $\mathbf{e}_{q_i}$ and $\mathbf{e}_{c_i^+}$ are the embeddings of the query q_i and its positive candidate c_i^+, respectively. The set $\mathcal{C}$ contains all in-batch candidates. Notably, both q and c can be images, text, or composed image-text data. The parameter τ modulates the penalties on negative samples and is set to 0.02 unless otherwise specified in this paper.

Methods	Backbone	# Params	CIRCO	CIRR		FashionIQ	GeneCIS
			mAP@5	R@1	R_s@1	R@10	R_s@1
SEARLE (Baldrati et al., 2023)	CLIP-B	165M	9.4	24.0	54.9	22.9	-
CIReVL (Karthik et al., 2023)	CLIP-B	12.3B[†]	14.9	23.9	60.2	28.3	15.9
LDRE (Yang et al., 2024)	CLIP-B	7.9B[†]	18.0	25.7	60.5	24.8	-
MagicLens-B (Zhang et al., 2024)	CLIP-B	166M	23.1	27.0	66.7	26.3	15.0
MagicLens-B[‡] (Zhang et al., 2024)	CoCa-B	267M	30.8	31.6	69.3	**35.2**	17.4[*]
MMRet-Base	CLIP-B	149M	**34.3**	**36.1**	**71.6**	31.9	**18.0**
Pic2Word (Saito et al., 2023)	CLIP-L	429M	8.7	23.9	-	24.7	11.2
PLI (Chen and Lai, 2023)	CLIP-L	428M	10.4	25.5	55.6	35.4	-
SEARLE (Baldrati et al., 2023)	CLIP-L	442M	11.7	24.2	53.8	25.6	12.3
CompoDiff (Gu et al., 2024a)	CLIP-L	568M	12.6	18.2	57.4	36.0	14.9
CIReVL (Karthik et al., 2023)	CLIP-L	12.5B[†]	18.6	24.6	59.5	28.6	15.9
LDRE (Yang et al., 2024)	CLIP-L	8.2B[†]	23.4	26.5	60.4	28.5	-
MagicLens-L (Zhang et al., 2024)	CLIP-L	465M	29.6	30.1	68.1	30.7	16.3
MagicLens-L[‡] (Zhang et al., 2024)	CoCa-L	613M	34.1[*]	33.3[*]	70.9[*]	**38.0**	16.7
MMRet-Large	CLIP-L	428M	**39.2**	**38.0**	**73.2**	34.6	**18.1**
LDRE (Yang et al., 2024)	CLIP-G	10.3B[†]	31.1	36.2	68.8	32.5	-
CIReVL (Karthik et al., 2023)	CLIP-G	14.6B[†]	26.8	34.7	68.0	32.2	17.4[*]
IP-CIR (Li et al., 2024c)	CLIP-G	43.8B[†]	32.8	39.3	70.0	**45.7**[*]	-
E5-V (Jiang et al., 2024a)	LLaVA-1.6	8.35B	19.1	33.9	-	31.8	-
MM-Embed (Lin et al., 2024)	LLaVA-1.6	7.57B	32.3	-	-	-	-
MMRet-MLLM	LLaVA-1.6	7.57B	**42.2**	**46.7**	**75.4**	35.6	**21.1**

Table 1: Zero-shot retrieval performance on various CIR benchmarks. [*] denotes the previous best performance for each benchmark prior to MMRet. [†] indicates methods with multiple components (e.g., GPT-3.5, Qwen1.5-32B); we report # parameters of components with known sizes. The CoCa-based MagicLens[‡] models are proprietary. Results in **bold** and underline denote the best and second-best performances for each model scale, respectively. Our MMRet model achieves state-of-the-art results across different model sizes and benchmarks, surpassing the previous SOTA by 8.1% on the main benchmark CIRCO, significantly advancing zero-shot CIR methods.

4 Experiments

In this section, we first evaluate the effectiveness of MegaPairs on zero-shot composed image retrieval (CIR) tasks in Section 4.1. Next, we explore the impact of MegaPairs on broader multimodal retrieval tasks in Section 4.2. Finally, we conduct detailed analysis on our MegaPairs in Section 4.3.

4.1 Zero-shot Performance on CIR tasks

4.1.1 Implementation Details

We utilize our MegaPairs dataset to perform multimodal contrastive training for our MMRet models. For the CLIP-based MMRet, we initialize the model using both the base[1] and large[2] versions of CLIP, referred to as MMRet-Base and MMRet-Large, respectively. For the MLLM-based MMRet, we leverage the LLaVA-1.6 Mistral 7B architecture[3] and initialize the model parameters accordingly, which we denote as MMRet-MLLM. The

training details of MMRet on MegaPairs can be found in Appendix B.

4.1.2 Benchmarks

We evaluate our MMRet in a zero-shot setting across four different composed image retrieval benchmarks: CIRCO (Baldrati et al., 2023), CIRR (Liu et al., 2021), FashionIQ (Wu et al., 2021), and GeneCIS (Vaze et al., 2023). Following previous practice (Zhang et al., 2024), CIRCO is considered our main benchmark due to its extensive candidate pool and high-quality annotations. Detailed information and metrics for each benchmark can be found in Appendix C.

4.1.3 Evaluation Results

The main evaluation results of MMRet across four benchmarks are shown in Table 1, with full results for each benchmark provided in Appendix D. We have identified three key observations:

(1) Our MMRet-MLLM model achieves leading performance across three of the four benchmarks. Specifically, on our main benchmark,

---

[1] https://huggingface.co/openai/clip-vit-base-patch16
[2] https://huggingface.co/openai/clip-vit-large-patch14
[3] https://huggingface.co/llava-hf/llava-v1.6-mistral-7b-hf

Models	Per Meta-Task Score				Overall
	Classification	VQA	Retrieval	Grounding	
number of datasets	10	10	12	4	36
BLIP2 (Li et al., 2023)	27.0	4.2	33.9	47.0	25.2
SigLIP (Zhai et al., 2023)	40.3	8.4	31.6	59.5	34.8
CLIP (Radford et al., 2021)	42.8	9.1	53.0	51.8	37.8
OpenCLIP (Cherti et al., 2023)	**47.8**	10.9	52.3	53.3	39.7
UniIR (Wei et al., 2024)	42.1	<u>15.0</u>	**60.1**†	**62.2**	<u>42.8</u>
MagicLens (Zhang et al., 2024)	38.8	8.3	35.4	26.0	27.8
E5-V (LLaVA-1.6) (Jiang et al., 2024a)	21.8	4.9	11.5	19.0	13.3
MMRet-MLLM (LLaVA-1.6)	<u>47.2</u>	**18.4**	<u>56.5</u>	**62.2**	**44.0**

Table 2: Zero-shot performance on the Massive Multimodal Embedding Benchmark (MMEB). †UniIR was trained on M-BEIR (Wei et al., 2024), which includes 10 of the 12 datasets in the MMEB retrieval tasks, it does not strictly adhere to a zero-shot setting. In contrast, our MMRet-MLLM, trained exclusively on the MegaPairs dataset, achieves state-of-the-art zero-shot performance in overall scores and multiple meta-tasks on MMEB.

CIRCO, MMRet-MLLM surpasses the current SOTA CoCa-based MagicLens-L by achieving 42.2% mAP@5 compared to 34.1% (an increase of 8.1%). On CIRR test set, it exceeds the current SOTA by 7.4% and 4.5% in R@1 and R_s@1, respectively. Additionally, on GeneCIS, it leads the current SOTA by 3.7% in R_s@1.

(2) MMRet exhibits superior performance across all model scales. For instance, MMRet-Base and MMRet-Large outperform comparable models by 4.5% and 4.7% in R@1 on the CIRR test set, respectively. Additionally, they surpass similar models by 3.5% and 5.1% in mAP@5 on the CIRCO benchmark. In the fashion-domain benchmark FashionIQ, while not achieving the highest scores, our CLIP-based MMRet shows competitive performance against other CLIP-based models.

(3) The MMRet-Base model surpasses most larger models, underscoring the exceptional quality of our MegaPairs dataset. Despite being our smallest model, MMRet-Base outperforms many larger models such as the MagicLens-L. For instance, it achieving the best result on CIRCO with a mAP@5 of 34.3%, excluding our own MMRet-Large and MMRet-MLLM models. It even exceeds the performance of models with dozens of times more parameters (e.g., MM-Embed), emphasizing the effectiveness of our MegaPairs dataset.

4.2 Performance on MMEB

To further validate the generalization ability of MegaPairs for broader multimodal embedding tasks, we evaluate MMRet on the Massive Mul-

timodal Embedding Benchmark (MMEB) (Jiang et al., 2024b). MMEB is a comprehensive benchmark that includes 36 datasets across four meta-task categories: *Classification*, *Visual Question Answering*, *Retrieval*, and *Visual Grounding*. It is designed to evaluate the quality of multimodal embeddings and assesses models across diverse combinations of text and image modalities. We present the performance of MMRet in both zero-shot and supervised fine-tuning scenarios. Following previous works (Jiang et al., 2024a,b), we conduct experiments using our MMRet-MLLM.

4.2.1 Zero-shot Performance

Implementation Details. In the zero-shot evaluation on MMEB, we directly utilize our MMRet-MLLM from Section 4.1, maintaining implementation details consistent with Section 4.1.1.

Metrics. We evaluate Precision@1 for all tasks, which measures the ratio of positive candidates ranked in the top position for all queries. We report the average scores for the four meta tasks as well as the overall average. Following the MMEB setting, we incorporate the predefined task-specific instructions into queries for all tasks during evaluation.

Results. The zero-shot performance of our MMRet-MLLM on MMEB is presented in Table 2. MMRet-MLLM achieved state-of-the-art zero-shot performance across various embedding meta-task, recording the highest overall average performance. Compared to the recent E5-V (Jiang et al., 2024a), which uses a similar LLaVA-1.6 (Liu

Models	Per Meta-Task Score				Average Score		
	Classification	VQA	Retrieval	Grounding	IND	OOD	Overall
number of datasets	10	10	12	4	20	16	36
CLIP	55.2	19.7	53.2	62.2	47.6	42.8	45.4
OpenCLIP	**56.0**	21.9	55.4	64.1	50.5	43.1	47.2
VLM2Vec (LLaVA-1.6)	54.7	50.3	56.2	64.0	61.0	47.5	55.0
VLM2Vec (Phi-3.5-V)	54.8	<u>54.9</u>	<u>62.3</u>	<u>79.5</u>	<u>66.5</u>	<u>52.0</u>	<u>60.1</u>
MMRet-MLLM	**56.0**	**57.4**	**69.9**	**83.6**	**68.0**	**59.1**	**64.1**

Table 3: Supervised fine-tuning results on the MMEB benchmark. The backbone of our MMRet-MLLM is LLaVA-1.6 (Liu et al., 2024). We compare our results with the following baselines: CLIP (Radford et al., 2021), OpenCLIP (Cherti et al., 2023), and two versions of VLM2Vec (Jiang et al., 2024b) that employ the LLaVA-1.6 (Liu et al., 2024) and Phi-3.5-V (Abdin et al., 2024) backbones. All baseline results are sourced from (Jiang et al., 2024b). IND: in-distribution dataset; OOD: out-of-distribution dataset.

et al., 2024) backbone for universal multimodal embedding, MMRet-MLLM trained on our MegaPairs dataset demonstrated superior performance. Notably, the second-best model, UniIR, was trained on M-BEIR (Wei et al., 2024), which encompasses datasets from 10 of the 12 retrieval meta-tasks in MMEB, and thus is not considered zero-shot for this meta-task. Consequently, our MLLM-Ret significantly outperforms the remaining methods in the retrieval meta-task and demonstrates strong generalization capabilities across all tasks.

4.2.2 Supervised Fine-tuning Performance

Implementation Details. We further fine-tune our MMRet-MLLM on MMEB to investigate the impact of MegaPairs on downstream task performance. The MMEB dataset includes 20 in-distribution (IND) datasets for training and 16 out-of-distribution (OOD) datasets for evaluation. We utilize the training sets from the 20 IND datasets, comprising approximately 662K data points. The learning rate is set to 5×10^{-6}, and we employ LoRA with a rank of 32. The batch size is set to 192, and we train for one epoch. Following the VLM2Vec configuration (Jiang et al., 2024b), we incorporate task-specific instructions into the queries during training.

Metrics. We employ the same metrics as outlined in Section 4.2.1. Additionally, we report the average scores for both the IND and OOD datasets.

Results. Table 3 compares the supervised fine-tuning performance of our MMRet model with various baselines on the MMEB dataset. Our MMRet-MLLM achieves state-of-the-art performance, with

an overall average Precision@1 of 64.1%. Compared to VLM2Vec (LLaVA-1.6) (Jiang et al., 2024b), which directly fine-tunes LLaVA-1.6 on MMEB, MMRet-MLLM enhances downstream task performance by 9.1% through multimodal contrastive training on our MegaPairs. Notably, our model shows improvements of 11.6% and 7.1% on out-of-distribution (OOD) datasets compared to the two versions of VLM2Vec, highlighting the superior generalization capability of our MegaPairs for broader downstream multimodal embedding tasks.

4.3 Detailed Investigation on MegaPairs

We first assess the quality and scalability of our MegaPairs dataset in Section 4.3.1. Next, we evaluate the effectiveness of the hard negative samples provided by MegaPairs in Section 4.3.2. Finally, we explore the strategies used for mining image pairs from open-domain image corpora in Section 4.3.3. Unless otherwise specified, all subsequent experiments are conducted using our MMRet-base model.

4.3.1 Data Scalability and Quality

We first evaluated the performance trend of MMRet by training it on different sizes of subsets from the MegaPairs dataset to verify its scalability. Subsequently, we compared it with existing datasets to highlight the high-quality features of MegaPairs.

Performance Scaling. As shown in Figure 2, the performance of MMRet-base across various benchmarks consistently improves with the increasing size of training data. This upward trend highlights the effectiveness and scalability of MegaPairs.

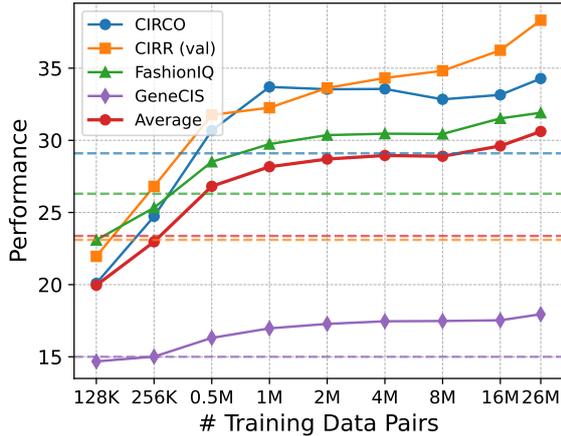**Dataset Quality Comparison with Exsiting**

Figure 2: Performance scaling of MMRet-base on the MegaPairs as data size increases. The dashed lines indicate the performance of MagicLens-B (CLIP) trained on their dataset of 36.7M data pairs.

Negatives		CIRCO	CIRR†	FIQ	CIS
Qry	*HN*	mAP@5	R@1	R@10	R$_s$@1
✗	✗	10.1	0.2	25.3	14.4
✓	✗	29.7	32.1	27.6	16.6
✓	✓	**32.3**	**33.7**	**30.1**	**17.0**

Table 4: Performance comparison of MMRet-base using different negative strategies at a 1M scale. *Qry*: query image negative; *HN*: our mined hard negatives. †We report CIRR validation set performance due to their test server submission limits.

Datasets. The dashed lines in Figure 2 represent the performance of the MagicLens-B (CLIP) model, trained on their 36.7M dataset (Zhang et al., 2024). Remarkably, with only 0.5M samples from our MegaPairs dataset, constituting less than 2% of MagicLens, MMRet significantly surpasses MagicLens across all benchmarks using the same CLIP-base backbone. This result underscores the superior quality and efficiency of our MegaPairs dataset.

4.3.2 The Impact of Hard Negatives

In MegaPairs, images from the retrieved set that are not the target are marked as hard negatives, providing a diverse and ample set of hard negative target images for each image pair. As shown in Table 4, compared to not using negatives or only using the query image as a negative, training with our mined hard negatives significantly enhances model performance across all benchmarks.

4.3.3 Data Pair Search Strategy

We explored the impact of various search strategies in constructing heterogeneous triplets. For a

Strategy			CIRCO	CIRR†	FIQ	CIS
D	*I*	*T*	mAP@5	R@1	R@10	R$_s$@1
✓	✗	✗	29.0	31.5	24.7	17.2
✗	✓	✗	30.0	30.0	29.6	15.3
✗	✗	✓	31.6	32.2	28.7	<u>17.3</u>
✓	✓	✗	31.0	32.1	28.5	17.1
✓	✗	✓	**32.4**	<u>33.3</u>	28.9	**17.5**
✗	✓	✓	32.2	<u>33.3</u>	<u>29.7</u>	16.4
✓	✓	✓	<u>32.3</u>	**33.7**	**30.1**	17.0

Table 5: Performance comparison of MMRet-base using different data pairing strategies at 1M scale. *D*: DINOv2 Encoder; *I*: CLIP Image Encoder; *T*: CLIP Text Encoder. FIQ and CIS represent the FashionIQ and GeneCIS benchmarks, respectively. † We report CIRR validation set performance due to test server submission limits.

fair comparison, we selected 1M data entries for each construction strategy and trained the model for 2000 steps.

Table 5 presents the results of various data pairing strategies across multiple benchmarks. Initially, when evaluating individual strategies, we observed that triplets based on text similarity achieved the highest zero-shot CIR performance. We hypothesize that text similarity captures more diverse relationships than image similarity. Furthermore, combining any two pairing strategies consistently outperformed using a single strategy. This enhancement is likely due to the increased diversity within the dataset, which is essential for training robust multimodal embedding models. Ultimately, employing all three strategies simultaneously provided the most robust performance across all datasets. As a result, this approach was chosen for constructing the MegaPairs.

5 Conclusion

In this paper, we introduce MegaPairs, a large-scale multimodal pairing dataset designed for training universal multimodal retrievers. MegaPairs comprises diverse image pairs from the open world, annotated with open-ended textual instructions that capture their visual and semantic relationships. Using MegaPairs, we trained our MMRet models, achieving state-of-the-art zero-shot performance in four composed image retrieval tasks and on the Massive Multimodal Embedding Benchmarks, which consists of 36 different datasets. Extensive experiments further demonstrate the generalization capability and high-quality features of MegaPairs.

Limitations

In constructing MegaPairs, we discovered that using diverse retrievers can generate richer image pairs. Our study employed three distinct retrievers, which offered substantial diversity. However, there remains potential to explore additional pairing methods, such as leveraging more advanced text domain retrievers (e.g., BGE (Xiao et al., 2024)) or incorporating varied strategies like image-text retrieval.

Ethics Statement

All images in our MegaPairs dataset are sourced from the Recap-Datacomp-1B dataset (Li et al., 2024b), and have undergone rigorous screening by the Datacomp team to remove harmful content (Gadre et al., 2024). Despite our best efforts, we acknowledge that these screenings may not be entirely comprehensive or without omissions. Additionally, we strongly discourage the use of MM-Ret models for encoding and retrieving sensitive content.

Acknowledgements

References

Marah Abdin, Jyoti Aneja, Hany Awadalla, Ahmed Awadallah, Ammar Ahmad Awan, Nguyen Bach, Amit Bahree, Arash Bakhtiari, Jianmin Bao, Harkirat Behl, et al. 2024. Phi-3 technical report: A highly capable language model locally on your phone. *arXiv preprint arXiv:2404.14219*.

Akari Asai, Timo Schick, Patrick Lewis, Xilun Chen, Gautier Izacard, Sebastian Riedel, Hannaneh Hajishirzi, and Wen-tau Yih. 2022. Task-aware retrieval with instructions. *arXiv preprint arXiv:2211.09260.*

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

Yingshan Chang, Mridu Narang, Hisami Suzuki, Guihong Cao, Jianfeng Gao, and Yonatan Bisk. 2022. Webqa: Multihop and multimodal qa. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16495–16504.

Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. 2021. Conceptual 12m: Pushing web-scale image-text pre-training to recognize long-tail visual concepts. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 3558–3568.

Jianlv Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. 2024a. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. *arXiv preprint arXiv:2402.03216.*

Junyang Chen and Hanjiang Lai. 2023. Pretrain like you inference: Masked tuning improves zero-shot composed image retrieval. *arXiv preprint arXiv:2311.07622.*

Wenhu Chen, Hexiang Hu, Xi Chen, Pat Verga, and William W. Cohen. 2022. Murag: Multimodal retrieval-augmented generator for open question answering over images and text. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 5558–5570. Association for Computational Linguistics.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. 2015. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325.*

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. 2024b. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821.*

Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. 2023. Reproducible scaling laws for contrastive language-image learning. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 2818–2829. IEEE.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. 2024. Scaling instruction-finetuned language models. *Journal of Machine Learning Research*, 25(70):1–53.

Niv Cohen, Rinon Gal, Eli A Meirom, Gal Chechik, and Yuval Atzmon. 2022. "this is my unicorn, fluffy": Personalizing frozen vision-language representations. In *European conference on computer vision*, pages 558–577. Springer.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. 2024. Datacomp: In search of the next generation of multimodal datasets. *Advances in Neural Information Processing Systems*, 36.

Geonmo Gu, Sanghyuk Chun, Wonjae Kim, HeeJae Jun, Yoohoon Kang, and Sangdoo Yun. 2024a. Compodiff: Versatile composed image retrieval with latent diffusion. *Trans. Mach. Learn. Res.*, 2024.

Geonmo Gu, Sanghyuk Chun, Wonjae Kim, Yoohoon Kang, and Sangdoo Yun. 2024b. Language-only efficient training of zero-shot composed image retrieval. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2024, Seattle, WA, USA, June 16-22, 2024*, pages 13225–13234. IEEE.

Sebastian Hofstätter, Sheng-Chieh Lin, Jheng-Hong Yang, Jimmy Lin, and Allan Hanbury. 2021. Efficiently teaching an effective dense retriever with balanced topic aware sampling. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 113–122.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net.

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR.

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

Ziyan Jiang, Rui Meng, Xinyi Yang, Semih Yavuz, Yingbo Zhou, and Wenhu Chen. 2024b. Vlm2vec: Training vision-language models for massive multimodal embedding tasks. *arXiv preprint arXiv:2410.05160*.

Shyamgopal Karthik, Karsten Roth, Massimiliano Mancini, and Zeynep Akata. 2023. Vision-by-language for training-free compositional image retrieval. *arXiv preprint arXiv:2310.09291*.

Chaofan Li, Zheng Liu, Shitao Xiao, Yingxia Shao, and Defu Lian. 2024a. Llama2vec: Unsupervised adaptation of large language models for dense retrieval. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3490–3500.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pretraining with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*.

Xianhang Li, Haoqin Tu, Mude Hui, Zeyu Wang, Bingchen Zhao, Junfei Xiao, Sucheng Ren, Jieru Mei, Qing Liu, Huangjie Zheng, et al. 2024b. What if we recaption billions of web images with llama-3? *arXiv preprint arXiv:2406.08478*.

You Li, Fan Ma, and Yi Yang. 2024c. Imagine and seek: Improving composed image retrieval with an imagined proxy. *arXiv preprint arXiv:2411.16752*.

Sheng-Chieh Lin, Chankyu Lee, Mohammad Shoeybi, Jimmy Lin, Bryan Catanzaro, and Wei Ping. 2024. Mm-embed: Universal multimodal retrieval with multimodal llms. *arXiv preprint arXiv:2411.02571*.

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

Zhenghao Liu, Chenyan Xiong, Yuanhuiyi Lv, Zhiyuan Liu, and Ge Yu. 2022. Universal vision-language dense retrieval: Learning a unified representation space for multi-modal retrieval. In *The Eleventh International Conference on Learning Representations*.

Zheyuan Liu, Cristian Rodriguez-Opazo, Damien Teney, and Stephen Gould. 2021. Image retrieval on real-life images with pre-trained vision-and-language models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2125–2134.

Man Luo, Zhiyuan Fang, Tejas Gokhale, Yezhou Yang, and Chitta Baral. 2023. 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), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 8573–8589. Association for Computational Linguistics.

Kenneth Marino, Mohammad Rastegari, Ali Farhadi, and Roozbeh Mottaghi. 2019. Ok-vqa: A visual question answering benchmark requiring external knowledge. In *Proceedings of the IEEE/cvf conference on computer vision and pattern recognition*, pages 3195–3204.

Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. 2021. Docvqa: A dataset for vqa on document images. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 2200–2209.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

Maxime Oquab, Timothée Darcet, Théo Moutakanni, Huy V. Vo, Marc Szafraniec, Vasil Khalidov, Pierre Fernandez, Daniel Haziza, Francisco Massa, Alaaeldin El-Nouby, Mido Assran, Nicolas Ballas, Wojciech Galuba, Russell Howes, Po-Yao Huang, Shang-Wen Li, Ishan Misra, Michael Rabbat, Vasu Sharma, Gabriel Synnaeve, Hu Xu, Hervé Jégou, Julien Mairal, Patrick Labatut, Armand Joulin, and Piotr Bojanowski. 2024. Dinov2: Learning robust visual features without supervision. *Trans. Mach. Learn. Res.*, 2024.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR.

Kuniaki Saito, Kihyuk Sohn, Xiang Zhang, Chun-Liang Li, Chen-Yu Lee, Kate Saenko, and Tomas Pfister. 2023. Pic2word: Mapping pictures to words for zeroshot composed image retrieval. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 19305–19314.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. 2022. LAION-5B: an open large-scale dataset for training next generation image-text models. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

Sahel Sharifymoghaddam, Shivani Upadhyay, Wenhu Chen, and Jimmy Lin. 2024. Unirag: Universal retrieval augmentation for multi-modal large language models. *arXiv preprint arXiv:2405.10311*.

Hongjin Su, Weijia Shi, Jungo Kasai, Yizhong Wang, Yushi Hu, Mari Ostendorf, Wen-tau Yih, Noah A Smith, Luke Zettlemoyer, and Tao Yu. 2022. One embedder, any task: Instruction-finetuned text embeddings. *arXiv preprint arXiv:2212.09741*.

Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. 2023. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389*.

Zhen Tan, Dawei Li, Song Wang, Alimohammad Beigi, Bohan Jiang, Amrita Bhattacharjee, Mansooreh Karami, Jundong Li, Lu Cheng, and Huan Liu. 2024. Large language models for data annotation and synthesis: A survey. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 930–957.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. Beir: A heterogenous benchmark for zero-shot evaluation of information retrieval models. *arXiv preprint arXiv:2104.08663*.

Sagar Vaze, Nicolas Carion, and Ishan Misra. 2023. Genecis: A benchmark for general conditional image similarity. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6862–6872.

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

Cong Wei, Yang Chen, Haonan Chen, Hexiang Hu, Ge Zhang, Jie Fu, Alan Ritter, and Wenhu Chen. 2024. Uniir: Training and benchmarking universal multimodal information retrievers. In *Computer Vision - ECCV 2024 - 18th European Conference, Milan, Italy, September 29-October 4, 2024, Proceedings, Part LXXXVII*, volume 15145 of *Lecture Notes in Computer Science*, pages 387–404. Springer.

Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. 2021. Finetuned language models are zero-shot learners. *arXiv preprint arXiv:2109.01652*.

Hui Wu, Yupeng Gao, Xiaoxiao Guo, Ziad Al-Halah, Steven Rennie, Kristen Grauman, and Rogerio Feris. 2021. Fashion iq: A new dataset towards retrieving images by natural language feedback. In *Proceedings of the IEEE/CVF Conference on computer vision and pattern recognition*, pages 11307–11317.

Shitao Xiao, Zheng Liu, Peitian Zhang, Niklas Muennighoff, Defu Lian, and Jian-Yun Nie. 2024. C-pack: Packed resources for general chinese embeddings. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '24, page 641–649, New York, NY, USA. Association for Computing Machinery.

Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul Bennett, Junaid Ahmed, and Arnold Overwijk. 2020. Approximate nearest neighbor negative contrastive learning for dense text retrieval. *arXiv preprint arXiv:2007.00808*.

Zhenyu Yang, Dizhan Xue, Shengsheng Qian, Weiming Dong, and Changsheng Xu. 2024. Ldre: Llm-based divergent reasoning and ensemble for zero-shot composed image retrieval. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 80–90.

Michihiro Yasunaga, Armen Aghajanyan, Weijia Shi, Richard James, Jure Leskovec, Percy Liang, Mike Lewis, Luke Zettlemoyer, and Wen-Tau Yih. 2023. Retrieval-augmented multimodal language modeling. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 39755–39769. PMLR.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.

Shi Yu, Chaoyue Tang, Bokai Xu, Junbo Cui, Junhao Ran, Yukun Yan, Zhenghao Liu, Shuo Wang, Xu Han, Zhiyuan Liu, et al. 2024. Visrag: Vision-based retrieval-augmented generation on multi-modality documents. *arXiv preprint arXiv:2410.10594*.

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 11975–11986.

Jianjin Zhang, Zheng Liu, Weihao Han, Shitao Xiao, Ruicheng Zheng, Yingxia Shao, Hao Sun, Hanqing Zhu, Premkumar Srinivasan, Weiwei Deng, et al. 2022. Uni-retriever: Towards learning the unified embedding based retriever in bing sponsored search. In *Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 4493–4501.

Kai Zhang, Yi Luan, Hexiang Hu, Kenton Lee, Siyuan Qiao, Wenhu Chen, Yu Su, and Ming-Wei Chang. 2024. Magiclens: Self-supervised image retrieval with open-ended instructions. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net.

Peitian Zhang, Shitao Xiao, Zheng Liu, Zhicheng Dou, and Jian-Yun Nie. 2023. Retrieve anything to augment large language models. *arXiv preprint arXiv:2310.07554*.

Junjie Zhou, Zheng Liu, Shitao Xiao, Bo Zhao, and Yongping Xiong. 2024. 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. Association for Computational Linguistics.

Appendix

A   Detailed Prompt for Annotating Open-Ended Instructions

To annotate open-ended instructions, we begin by using the MLLM to generate a detailed description of the commonalities and differences between the query image and the target image, where the corresponding prompt is illustrated in Figure 3. Subsequently, the description is refined by the LLM to produce textual instructions, with the associated prompt provided in Figure 4.

Source Image:

Source Image

Target Image:

Target Image

Carefully **observe the connections and differences** between Source Image and Target Image. Summarize the key commonality and detailed differences between them in under **WORD_NUM** words. Use this template: 'Both images are / have / exhibit / show [COMMON POINTS]. However, the Source Image [DIFFERENCES], whereas the Target Image [DIFFERENCES].'

Figure 3: The specific prompts for MLLM. The value of WORD_NUM ranges from 60 to 100 in our practical data generation to enhance the diversity of the generated description.

B   Training Details of MMRet on MegaPairs

For the CLIP-based MMRet, the training process employs a batch size of 2048, with each query paired with one positive image and four hard negatives. All input images are resized to 224x224 to match the model's configuration. During training, all CLIP parameters remain unfrozen. MMRet-base is trained for 15,000 steps, while MMRet-large is trained for 25,000 steps on the MegaPairs dataset.

For the MLLM-based MMRet, we use a batch size of 144 during training, with each query associated with one positive image and three hard neg-

Instruction: Based on the provided connection of two different images, **create interesting text queries** that can be used with the source image to retrieve the target image.

1) Replace similarities with the source image **using non-specific pronouns** to avoid revealing explicit details about the source image, and keep it concise.

2) **Detail the unique differences** that are present only in the target image.

Demonstrations:
- Connection: Both images show a hand holding another hand. However, the Source Image includes a small heart symbol, whereas the Target Image does not.
- Text Query: ["Look for the same interaction but devoid of any heart symbol.", "Remove the small heart symbol", "An Image of the same gesture, but without the small heart symbol."]

- Connection: Both images exhibit a small, white, propeller-driven aircraft. However, the Source Image is parked outdoors with a clear sky in the background, whereas the Target Image is indoors in a hangar with artificial lighting.
- Text Query: ["Find a similar vehicle indoors in a hangar with artificial lighting.", "Similar aircraft located indoors within a hangar and under artificial lighting.", "What if this plane is put indoors?"]
......

Your Task:

- Connection: Descriptions from MLLM

- Text Query:

Figure 4: The specific prompts for LLM. The figure showcases two demonstrations, while in our practical data generation process, five demonstrations are randomly selected from a pool of 50 and fed into the LLM.

atives. We apply LoRA (Hu et al., 2022) to both the ViT encoder and the LLM backbone of LLaVA-1.6, setting the LoRA rank to 32. Although the original model supports variable resolution image processing, we use a fixed resolution of 512x512 for all images to manage the token sequence length. MMRet-MLLM is trained for 20,000 steps on the MegaPairs dataset.

For both CLIP-based and MLLM-based MMRet models, we set an initial learning rate of 5×10^{-6} and employ a linear decay strategy.

C Detailed Information and Evaluation Metrics of Zero-Shot CIR Benchmarks

The detailed information and metrics of our evaluation in zero-shot composed image retrieval (CIR) tasks for each benchmark are as follows:

CIRCO (Baldrati et al., 2023) is a challenging zero-shot CIR benchmark comprising 123,403 candidate natural images. We evaluete our MMRet models on its test set, which contains 800 composed image-text queries, each annotated with mul-

tiple ground-truth images. We use mean Average Precision (mAP) as the evaluation metric. Due to its extensive candidate pool and high-quality annotations, CIRCO serves as a robust and comprehensive benchmark for zero-shot CIR evaluation, and we consider it our main benchmark.

CIRR (Liu et al., 2021) is the first dataset for conducting the CIR task using natural images. We conduct zero-shot evaluations on its test set, which comprises 4,148 queries and a corpus of 2,315 images. Each query in CIRR is annotated with exactly one positive target image, but it suffers from some false negative issues. For each query, CIRR provides a subset retrieval setting that retrieves target images from a small corpus. We assess both standard and subset retrieval performance using recall metrics (R and R_s).

FashionIQ (Wu et al., 2021) is another CIR task focusing on fashion products. We conduct zero-shot evaluations on its validation set, which includes 6,016 queries and 15,536 images. FashionIQ comprises three sub-tasks: dress, shirt, and toptee. We

19088

evaluate each sub-task separately and report their average recall values.

GeneCIS (Vaze et al., 2023) is a benchmark for conditional image similarity measurement, comprising four sub-tasks about changing or focusing the attribute or object in the given image. In each sub-task, models need to retrieve the most similar images from a dedicated small subset for the given query image and the condition keyword. We approach it as a CIR task by combining the query image and the text description of the sub-task derived from the condition keyword as a composed image-text query. Each query's candidate subset averages 13.8 images, and the mean R_s across all four subsets is reported.

D Full results on CIR Benchmarks

We report the full results on four CIR benchmarks (Baldrati et al., 2023; Liu et al., 2021; Wu et al., 2021; Vaze et al., 2023) in Tables 6, 7, 8, and 9, respectively. Our MMRet model achieves state-of-the-art performance across various model sizes on the CIRCO, CIRR, and GeneCIS benchmarks.

E Full Results on MMEB Benchmark

We list the full results on the MMEB benchmark (Jiang et al., 2024b) in Table 10. The MMEB benchmark consists of 36 datasets spanning four meta-task categories, including 20 in-distribution datasets and 16 out-of-distribution (OOD) datasets. The results on the OOD datasets are highlighted with a gray background in the table. Our MMRet model achieves state-of-the-art performance in both zero-shot and fine-tuning settings. Notably, MMRet surpasses the second-best performance on the OOD datasets, demonstrating its remarkable generalization capability.

F Visualized Examples of MegaPairs

We present several examples of MegaPairs in Figure 5. Each row corresponds to a single example, where the query item, comprising an image and its corresponding alt-text caption, is associated with multiple target images. These target images include both visually similar ones and semantically related ones beyond visual features.

For example, in the 4th row, the query image showing an ottoman with the alt-text caption Round ottoman, tufted surface is paired with target items that feature visually similar images (e.g., the 1st target image, which shows an ottoman, and the 3rd target image, which depicts a sofa with a similar style) as well as semantically related images that transcend visual features (e.g., the 2nd and 4th target images, depicting the interior of a car and a living room wall, respectively. These share few visual features with the query image but also exhibit a tufted surface). In the 5th row, the query image showing an F1 car with the alt-text caption AMG F1 W09 is paired with target items featuring visually similar images (e.g., the 1st target image, which shows an F1 car in red, and the 3rd target image, which displays a race scene with multiple F1 cars) as well as semantically related images that transcend visual features (e.g., the 2nd target image, which shows an F1 driver, and the 4th target image, depicting an F1 circuit. These images bear no visual similarity to the query image but share the F1 concept).

G Qualitative Results of MMRet on Zero-shot CIR Tasks

We present several top-5 retrieved images of our MMRet and the SOTA MagicLens (Zhang et al., 2024) on zero-shot CIR tasks, as shown in Figure 6. Since only the CLIP-based checkpoint is available for MagicLens, we select the CLIP-L backbone for both methods. 1) For the blue ties query, MMRet accurately interprets the query and identifies both the specific attire and indoor setting, retrieving multiple images that meet the specified requirements. In contrast, MagicLens focuses solely on the individual object, overlooking the broader semantic context. 2) For the sweet, beverage, boats and sky query, MMRet demonstrates a solid understanding of real-world entity concepts, successfully integrating both foreground and background elements to retrieve the most relevant image. 3) The success on the bench top query highlights MMRet's ability to comprehend specific pose and angle requirements. 4) The success on the darker ground and closer distance query illustrates MMRet's capacity to recognize lighting conditions and shooting distance. 5) The success on the whell in the air query indicates that MMRet can identify dynamic actions and contextual scene elements.

Methods	Backbone	# Params	mAP@5	mAP@10	mAP@25	mAP@50
PALAVRA (Cohen et al., 2022)	CLIP-B	176M	4.6	5.3	6.3	6.8
PLI (Chen and Lai, 2023)	BLIP-B	224M	7.1	8.0	9.2	9.7
SEARLE (Baldrati et al., 2023)	CLIP-B	165M	9.4	9.9	11.1	11.8
CIReVL (Karthik et al., 2023)	CLIP-B	12.3B[†]	14.9	15.4	17.0	17.8
LDRE (Yang et al., 2024)	CLIP-B	7.9B[†]	18.0	18.3	20.2	21.1
MagicLens-B (Zhang et al., 2024)	CLIP-B	166M	23.1	23.8	25.8	26.7
MagicLens-B[‡] (Zhang et al., 2024)	CoCa-B	267M	30.8	32.0	34.5	35.6
MMRet-Base	CLIP-B	149M	**34.3**	**35.0**	**37.6**	**38.7**
Pic2Word (Saito et al., 2023)	CLIP-L	429M	8.7	9.5	10.6	11.3
PLI (Chen and Lai, 2023)	CLIP-L	428M	10.4	11.6	13.0	13.7
SEARLE (Baldrati et al., 2023)	CLIP-L	442M	11.7	12.7	14.3	15.1
CIReVL (Karthik et al., 2023)	CLIP-L	12.5B[†]	18.6	19.0	20.9	21.8
LinCIR (Gu et al., 2024b)	CLIP-L	442M	12.6	13.6	15.0	15.9
CompoDiff (Gu et al., 2024a)	CLIP-L	568M	12.6	13.4	15.8	16.4
MagicLens-L (Zhang et al., 2024)	CLIP-L	465M	29.6	30.8	33.4	34.4
MagicLens-L[‡] (Zhang et al., 2024)	CoCa-L	613M	34.1	35.4	38.1	39.2
MMRet-Large	CLIP-L	428M	**39.2**	**40.2**	**42.9**	**44.0**
Pic2Word (Saito et al., 2023)	CLIP-H	987M	11.7	12.3	13.7	14.4
SEARLE (Baldrati et al., 2023)	CLIP-H	1.0B	16.1	16.9	18.8	19.7
LinCIR (Gu et al., 2024b)	CLIP-H	1.0B	17.6	18.5	20.5	21.4
Pic2Word (Saito et al., 2023)	CLIP-G	2.5B	5.5	5.6	6.7	7.1
SEARLE (Baldrati et al., 2023)	CLIP-G	2.6B	13.2	13.9	15.3	16.0
CompoDiff (Gu et al., 2024a)	CLIP-G	2.9B	15.3	17.7	19.4	-
CIReVL (Karthik et al., 2023)	CLIP-G	14.6B[†]	26.8	27.6	30.0	31.0
LinCIR (Gu et al., 2024b)	CLIP-G	2.6B	19.7	21.0	23.1	24.2
LDRE (Yang et al., 2024)	CLIP-G	10.3B[†]	31.1	32.2	35.0	36.0
IP-CIR (Li et al., 2024c)	CLIP-G	43.8B[†]	32.8	34.3	36.9	38.0
E5-V (Jiang et al., 2024a)	LLaVA-1.6	8.35B	19.1	-	-	-
MM-Emded (Lin et al., 2024)	LLaVA-1.6	7.57B	32.3	-	-	-
MMRet-MLLM	LLaVA-1.6	7.57B	**42.2**	**43.4**	**46.5**	**47.6**

Table 6: Full results on the CIRCO benchmark (Baldrati et al., 2023). [†] indicates methods with multiple components (e.g., GPT-3.5, Qwen1.5-32B); we report # parameters of components with known sizes. The CoCa-based MagicLens[‡] models are proprietary. Results in **bold** denote the best performances for each model scale.

Methods	Backbone	# Params	Index Set				Subset Set		
			R@1	R@5	R@10	R@50	R@1	R@2	R@3
PALAVRA (Cohen et al., 2022)	CLIP-B	176M	16.6	43.5	58.5	84.0	41.6	65.3	80.9
PLI (Chen and Lai, 2023)	BLIP-B	224M	27.2	58.9	71.4	91.3	55.1	77.4	89.1
SEARLE (Baldrati et al., 2023)	CLIP-B	165M	24.0	53.4	66.8	89.8	54.9	76.6	88.2
CIReVL (Karthik et al., 2023)	CLIP-B	12.3B[†]	23.9	52.5	66.0	87.0	60.2	80.1	90.2
LDRE (Yang et al., 2024)	CLIP-B	7.9B[†]	25.7	55.1	69.0	89.9	60.5	80.7	90.7
MagicLens-B (Zhang et al., 2024)	CLIP-B	166M	27.0	58.0	70.9	91.1	66.7	83.9	92.4
MagicLens-B[‡] (Zhang et al., 2024)	CoCa-B	267M	31.6	64.0	76.9	93.8	69.3	86.0	94.0
MMRet-Base	CLIP-B	149M	**36.1**	**68.1**	**79.5**	**94.7**	**71.6**	**87.2**	**94.0**
Pic2Word (Saito et al., 2023)	CLIP-L	429M	23.9	51.7	65.3	87.8	-	-	-
PLI (Chen and Lai, 2023)	CLIP-L	428M	25.5	54.6	67.6	88.7	55.6	77.5	89.5
SEARLE (Baldrati et al., 2023)	CLIP-L	442M	24.2	52.5	66.3	88.8	53.8	75.0	88.2
CIReVL (Karthik et al., 2023)	CLIP-L	12.5B[†]	24.6	52.3	64.9	86.3	59.5	79.9	89.7
LinCIR (Gu et al., 2024b)	CLIP-L	442M	25.0	53.3	66.7	-	57.1	77.4	88.9
CompoDiff (Gu et al., 2024a)	CLIP-L	568M	18.2	53.1	70.8	90.3	57.4	77.1	87.9
MagicLens-L (Zhang et al., 2024)	CLIP-L	465M	30.1	61.7	74.4	92.6	68.1	84.8	93.2
MagicLens-L[‡] (Zhang et al., 2024)	CoCa-L	613M	33.3	67.0	77.9	94.4	70.9	87.3	94.5
MMRet-Large	CLIP-L	428M	**38.0**	**70.3**	**81.1**	**94.7**	**73.2**	**88.0**	94.3
Pic2Word (Saito et al., 2023)	CLIP-H	987M	32.9	63.1	73.9	-	62.2	81.4	91.2
SEARLE (Baldrati et al., 2023)	CLIP-H	1.0B	34.0	64.0	75.3	-	64.6	83.2	92.8
LinCIR (Gu et al., 2024b)	CLIP-H	1.0B	33.8	63.5	73.4	-	62.4	81.5	92.1
Pic2Word (Saito et al., 2023)	CLIP-G	2.5B	30.4	58.1	69.2	-	68.9	85.5	93.0
SEARLE (Baldrati et al., 2023)	CLIP-G	2.6B	34.8	64.1	75.1	-	68.7	84.7	93.2
CompoDiff (Gu et al., 2024a)	CLIP-G	2.9B	26.7	55.1	74.5	92.0	64.5	82.4	91.8
CIReVL (Karthik et al., 2023)	CLIP-G	14.6B[†]	34.7	64.3	75.1	91.7	68.0	84.9	93.2
LinCIR (Gu et al., 2024b)	CLIP-G	2.6B	35.3	64.7	76.1	-	63.4	82.2	92.0
LDRE (Yang et al., 2024)	CLIP-G	10.3B[†]	36.2	66.4	77.3	94.0	68.8	85.7	93.8
IP-CIR (Li et al., 2024c)	CLIP-G	43.8B[†]	39.3	70.1	80.0	94.9	70.0	86.9	94.2
E5-V (Jiang et al., 2024a)	LLaVA-1.6	8.35B	33.9	64.1	75.9	93.5	-	-	-
MMRet-MLLM	LLaVA-1.6	7.57B	**46.7**	**76.0**	**85.1**	**96.5**	**75.4**	**89.6**	**95.7**

Table 7: Full results on the CIRR benchmark (Liu et al., 2021). [†] indicates methods with multiple components (e.g., GPT-3.5, Qwen1.5-32B); we report # parameters of components with known sizes. The CoCa-based MagicLens[‡] models are proprietary. Results in **bold** denote the best performance for each model scale.

Methods	Backbone	# Params	Dress		Shirt		Toptee		Overall	
			R@10	R@50	R@10	R@50	R@10	R@50	R@10	R@50
PALAVRA (Cohen et al., 2022)	CLIP-B	176M	17.3	35.9	21.5	37.1	20.6	38.8	19.8	37.3
PLI (Chen and Lai, 2023)	BLIP-B	224M	28.6	50.8	38.1	57.8	40.9	62.7	**35.9**	57.1
SEARLE (Baldrati et al., 2023)	CLIP-B	165M	18.5	39.5	24.4	41.6	25.7	46.5	22.9	42.5
CIReVL (Karthik et al., 2023)	CLIP-B	12.3B[†]	25.3	46.4	28.4	47.8	31.2	53.9	28.3	49.4
LDRE (Yang et al., 2024)	CLIP-B	7.9B[†]	27.4	46.3	20.0	41.8	27.1	48.8	24.8	45.6
MagicLens-B (Zhang et al., 2024)	CLIP-B	166M	21.5	41.3	27.3	48.8	30.2	52.3	26.3	47.4
MagicLens-B[‡] (Zhang et al., 2024)	CoCa-B	267M	29.0	48.9	36.5	55.5	40.2	61.9	35.2	55.4
MMRet-Base	CLIP-B	149M	26.1	49.3	33.7	53.4	36.0	57.5	31.9	53.4
Pic2Word (Saito et al., 2023)	CLIP-L	429M	20.0	40.2	26.2	43.6	27.9	47.4	24.7	43.7
PLI (Chen and Lai, 2023)	CLIP-L	428M	28.1	51.1	38.6	58.5	39.4	62.7	35.4	57.4
SEARLE (Baldrati et al., 2023)	CLIP-L	442M	20.5	43.1	26.9	45.6	29.3	50.0	25.6	46.2
CIReVL (Karthik et al., 2023)	CLIP-L	12.5B[†]	24.8	44.8	29.5	47.4	31.4	53.7	28.6	48.6
LinCIR (Gu et al., 2024b)	CLIP-L	442M	20.9	42.4	29.1	46.8	28.8	50.2	26.3	46.5
CompoDiff (Gu et al., 2024a)	CLIP-L	568M	32.2	46.3	37.7	49.1	38.1	50.6	36.0	48.6
MagicLens-L (Zhang et al., 2024)	CLIP-L	465M	25.5	46.1	32.7	53.8	34.0	57.7	30.7	52.5
MagicLens-L[‡] (Zhang et al., 2024)	CoCa-L	613M	32.3	52.7	40.5	59.2	41.4	63.0	38.0	58.2
MMRet-Large	CLIP-L	428M	29.7	50.3	37.0	56.1	37.0	59.3	34.6	55.2
Pic2Word (Saito et al., 2023)	CLIP-H	987M	28.0	51.5	36.9	56.0	40.2	62.0	35.0	56.5
SEARLE (Baldrati et al., 2023)	CLIP-H	1.0B	28.5	51.1	36.5	55.5	38.8	60.9	34.6	55.8
LinCIR (Gu et al., 2024b)	CLIP-H	1.0B	29.8	52.1	36.9	57.8	42.1	62.5	36.3	57.5
Pic2Word (Saito et al., 2023)	CLIP-G	2.5B	25.4	47.7	33.2	50.4	35.2	57.6	31.3	51.9
SEARLE (Baldrati et al., 2023)	CLIP-G	2.6B	28.2	50.3	36.5	55.4	39.8	61.5	34.8	55.7
CompoDiff (Gu et al., 2024a)	CLIP-G	2.9B	37.8	49.1	41.3	55.2	44.3	56.4	39.0	51.7
CIReVL (Karthik et al., 2023)	CLIP-G	14.6B[†]	27.1	49.5	33.7	51.4	35.8	56.1	32.2	52.4
LinCIR (Gu et al., 2024b)	CLIP-G	2.6B	38.1	60.9	46.8	65.1	50.5	71.1	45.1	65.7
LDRE (Yang et al., 2024)	CLIP-G	10.3B[†]	35.9	58.6	26.1	51.1	35.4	56.7	32.5	55.5
IP-CIR (Li et al., 2024c)	CLIP-G	43.8B[†]	48.0	66.7	39.0	61.0	50.2	71.1	**45.7**	66.3
E5-V (Jiang et al., 2024a)	LLaVA-1.6	8.35B	36.4	56.4	23.8	47.5	35.3	57.5	31.8	53.8
MMRet-MLLM	LLaVA-1.6	7.57B	29.1	50.5	38.5	58.6	39.2	60.6	35.6	56.6

Table 8: Full results on the FashionIQ benchmark (Wu et al., 2021). [†] indicates methods with multiple components (e.g., GPT-3.5, Qwen1.5-32B); we report # parameters of components with known sizes. The CoCa-based MagicLens[‡] models are proprietary. Results in **bold** denote the best performance for each model scale.

Methods	Backbone	# Params	Focus Attribute			Change Attribute			Focus Object			Change Object			Avg
			R@1	R@2	R@3	R@1	R@2	R@3	R@1	R@2	R@3	R@1	R@2	R@3	R@1
CIReVL (Karthik et al., 2023)	CLIP-B	12.3B[†]	17.9	29.4	40.4	14.8	25.8	35.8	14.6	24.3	33.3	16.1	27.8	37.6	15.9
MagicLens-B (Zhang et al., 2024)	CLIP-B	166M	15.5	28.4	39.1	12.3	23.0	32.1	14.4	26.2	35.5	17.7	28.4	39.2	15.0
MagicLens-B[‡] (Zhang et al., 2024)	CoCa-B	267M	16.2	27.8	38.6	16.2	27.2	36.6	17.1	27.7	38.2	20.2	32.2	42.9	17.4
MMRet-Base	CLIP-B	149M	18.3	30.9	39.6	15.2	25.6	34.8	16.6	27.3	35.8	21.7	34.9	45.0	**18.0**
Pic2Word (Saito et al., 2023)	CLIP-L	429M	15.7	28.2	38.7	13.9	24.7	33.1	8.4	18.0	25.8	6.7	15.1	24.0	11.2
SEARLE (Baldrati et al., 2023)	CLIP-L	442M	17.0	29.7	40.7	16.4	25.3	34.1	8.0	16.9	25.6	7.9	16.8	24.8	12.3
CIReVL (Karthik et al., 2023)	CLIP-L	12.5B[†]	19.5	31.8	42.0	14.4	26.0	35.2	12.3	21.8	30.5	17.2	28.9	37.6	15.9
LinCIR (Gu et al., 2024b)	CLIP-L	442M	16.9	30.0	41.5	16.2	28.0	36.8	8.3	17.4	26.2	7.4	15.7	25.0	12.2
CompoDiff (Gu et al., 2024a)	CLIP-L	568M	13.5	24.3	36.1	19.2	28.6	37.2	8.1	16.4	25.1	18.7	31.7	40.6	14.9
MagicLens-L (Zhang et al., 2024)	CLIP-L	465M	16.1	28.7	39.0	15.6	27.5	36.3	16.3	26.2	35.5	17.1	29.5	39.7	16.3
MagicLens-L[‡] (Zhang et al., 2024)	CoCa-L	613M	16.6	28.7	39.3	16.0	27.5	36.5	15.7	27.6	37.3	18.7	31.7	40.2	16.7
MMRet-Large	CLIP-L	428M	18.4	30.0	38.5	15.4	27.6	35.7	17.4	26.6	36.3	21.0	34.0	42.4	**18.1**
Pic2Word (Saito et al., 2023)	CLIP-H	987M	18.6	30.7	42.1	13.2	23.9	33.1	9.2	17.6	27.1	6.6	16.5	25.4	11.9
SEARLE (Baldrati et al., 2023)	CLIP-H	1.0B	18.8	31.5	42.3	15.5	26.9	35.9	10.6	18.7	26.5	8.5	17.9	26.2	13.3
LinCIR (Gu et al., 2024b)	CLIP-H	1.0B	19.6	31.5	41.6	16.6	27.6	37.5	9.8	18.8	27.9	9.0	17.6	25.7	13.8
Pic2Word (Saito et al., 2023)	CLIP-G	2.5B	12.5	23.4	33.7	11.7	21.9	30.9	9.9	19.3	27.4	8.6	18.2	26.1	10.7
SEARLE (Baldrati et al., 2023)	CLIP-G	2.6B	16.3	29.4	40.7	16.2	27.3	35.5	10.8	18.2	27.9	8.3	15.6	25.8	12.9
CompoDiff (Gu et al., 2024a)	CLIP-G	2.9B	14.3	26.7	38.4	19.7	28.8	37.4	9.2	19.1	25.8	18.7	31.7	40.2	15.5
CIReVL (Karthik et al., 2023)	CLIP-G	14.6B[†]	20.5	34.0	44.5	16.1	28.6	39.4	14.7	25.2	33.0	18.1	31.2	41.0	17.4
LinCIR (Gu et al., 2024b)	CLIP-G	2.6B	19.1	33.0	42.3	17.6	30.2	38.1	10.1	19.1	28.1	7.9	16.3	25.7	13.7
MMRet-MLLM	LLaVA-1.6	7.57B	18.4	31.4	41.0	16.7	27.7	36.4	22.4	32.5	41.6	26.9	40.4	49.9	**21.1**

Table 9: Full results on the GeneCIS benchmark (Vaze et al., 2023). [†] indicates methods with multiple components (e.g., GPT-3.5, Qwen1.5-32B); we report # parameters of components with known sizes. The CoCa-based MagicLens[‡] models are proprietary. Results in **bold** denote the best performance for each model scale.

Task	Zero-shot								Fine-Tune	
	CLIP	OpenCLIP	SigLIP	BLIP2	MagicLens	E5-V	UniIR	MMRet	VLM2Vec	MMRet
Classification (10 tasks)										
ImageNet-1K	55.8	63.5	45.4	10.3	48.0	9.6	53.7	49.1	65.6	58.8
N24News	34.7	38.6	13.9	36.0	33.7	23.4	33.9	45.8	79.5	71.3
HatefulMemes	51.1	51.7	47.2	49.6	49.0	49.7	51.0	51.0	67.1	53.7
VOC2007	50.7	52.4	64.3	52.1	51.6	49.9	62.7	74.6	88.6	85.0
SUN397	43.4	68.8	39.6	34.5	57.0	33.1	61.7	60.1	72.7	70.0
Place365	28.5	37.8	20.0	21.5	31.5	8.6	38.0	35.3	42.6	43.0
ImageNet-A	25.5	14.2	42.6	3.2	8.0	2.0	12.9	31.6	19.3	36.1
ImageNet-R	75.6	83.0	75.0	39.7	70.9	30.8	61.6	66.2	70.2	71.6
ObjectNet	43.4	51.4	40.3	20.6	31.6	7.5	37.1	49.2	29.5	55.8
Country-211	19.2	16.8	14.2	2.5	6.2	3.1	8.8	9.3	13.0	14.7
All Classification	42.8	47.8	40.3	27.0	38.8	21.8	42.1	47.2	54.8	56.0
VQA (10 tasks)										
OK-VQA	7.5	11.5	2.4	8.7	12.7	8.9	24.5	28.0	63.2	73.3
A-OKVQA	3.8	3.3	1.5	3.2	2.9	5.9	10.6	11.6	50.2	56.7
DocVQA	4.0	5.3	4.2	2.6	3.0	1.7	5.6	12.6	78.4	78.5
InfographicsVQA	4.6	4.6	2.7	2.0	5.9	2.3	5.0	10.6	40.8	39.3
ChartQA	1.4	1.5	3.0	0.5	0.9	2.4	1.8	2.4	59.0	41.7
Visual7W	4.0	2.6	1.2	1.3	2.5	5.8	12.3	9.0	47.7	49.5
ScienceQA	9.4	10.2	7.9	6.8	5.2	3.6	11.6	23.3	43.4	45.2
VizWiz	8.2	6.6	2.3	4.0	1.7	2.6	19.2	25.9	39.2	51.7
GQA	41.3	52.5	57.5	9.7	43.5	7.8	49.3	41.3	60.7	59.0
TextVQA	7.0	10.9	1.0	3.3	4.6	3.2	10.6	18.9	66.1	79.0
All VQA	9.1	10.9	8.4	4.2	8.3	4.9	15.0	18.4	54.9	57.4
Retrieval (12 tasks)										
VisDial	30.7	25.4	21.5	18.0	24.8	9.2	37.6	62.6	73.3	83.0
CIRR	12.6	15.4	15.1	9.8	39.1	6.1	53.2	65.7	47.8	61.4
VisualNews_t2i	78.9	74.0	51.0	48.1	50.7	13.5	63.6	45.7	67.2	74.2
VisualNews_i2t	79.6	78.0	52.4	13.5	21.1	8.1	68.8	53.4	70.7	78.1
MSCOCO_t2i	59.5	63.6	58.3	53.7	54.1	20.7	72.0	68.7	70.6	78.6
MSCOCO_i2t	57.7	62.1	55.0	20.3	40.0	14.0	74.1	56.7	66.5	72.4
NIGHTS	60.4	66.1	62.9	56.5	58.1	4.2	69.7	59.4	66.1	68.3
WebQA	67.5	62.1	58.1	55.4	43.0	17.7	86.3	76.3	88.1	90.2
FashionIQ	11.4	13.8	20.1	9.3	11.2	2.8	39.3	31.5	12.9	54.9
Wiki-SS-NQ	55.0	44.6	55.1	28.7	18.7	8.6	11.3	25.4	56.6	24.9
OVEN	41.1	45.0	56.0	39.5	1.6	5.9	66.6	73.0	47.3	87.5
EDIS	81.0	77.5	23.6	54.4	62.6	26.8	78.2	59.9	79.9	65.6
All Retrieval	53.0	52.3	31.6	33.9	35.4	11.5	60.1	56.5	62.3	69.9
Visual Grounding (4 tasks)										
MSCOCO	33.8	34.5	46.4	28.9	22.1	10.8	46.6	42.7	67.3	76.8
RefCOCO	56.9	54.2	70.8	47.4	22.8	11.9	67.8	69.3	84.7	89.8
RefCOCO-matching	61.3	68.3	50.8	59.5	35.6	38.9	62.9	63.2	79.2	90.6
Visual7W-pointing	55.1	56.3	70.1	52.0	23.4	14.3	71.3	73.5	86.8	77.0
All Visual Grounding	51.8	53.3	59.5	47.0	26.0	19.0	62.2	62.2	79.5	83.6
Final Score (36 tasks)										
All	37.8	39.7	34.8	25.2	27.8	13.3	42.8	44.0	60.1	64.1
All IND	37.1	39.3	32.3	25.3	31.0	14.9	44.7	43.5	66.5	59.1
All OOD	38.7	40.2	38.0	25.1	23.7	11.5	40.4	44.3	52.0	68.0

Table 10: The detailed results on the MMEB benchmark (Jiang et al., 2024b). We report the performance of our MMRet under both zero-shot and fine-tuning settings.

Figure 5: The visualized examples of MegaPairs. Each row represents a single example, with the query item highlighted in a blue rectangle and the target items enclosed within a dashed box.
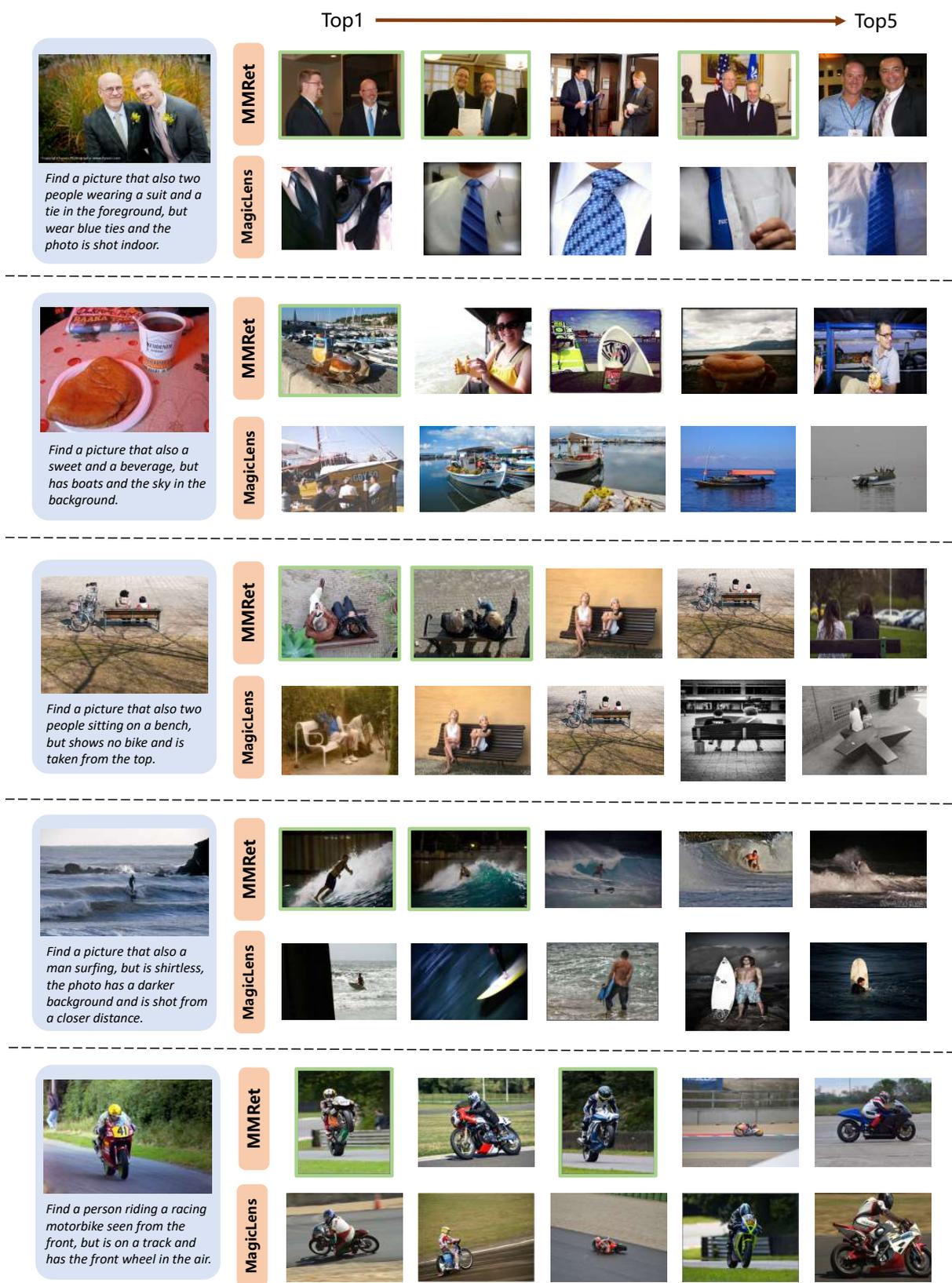
Top1 ⟶ Top5

Figure 6: Top-5 retrieved images of MMRet and MagicLens on zero-shot CIR tasks, both using the CLIP-L backbone. Queries are shown with a blue background, and the most correct retrieved images are marked with green outlines.