UniIR: Training and Benchmarking Universal Multimodal Information Retrievers

Cong Wei[†], **Yang Chen**[‡], **Haonan Chen**[†], **Hexiang Hu**[◦], **Ge Zhang**[†], **Jie Fu**[§], **Alan Ritter**[‡], **Wenhu Chen**[†]

[†]University of Waterloo [‡]Georgia Institute of Technology [§]Hong Kong University of Science and Technology [◦]Google DeepMind

https://tiger-ai-lab.github.io/UniIR/

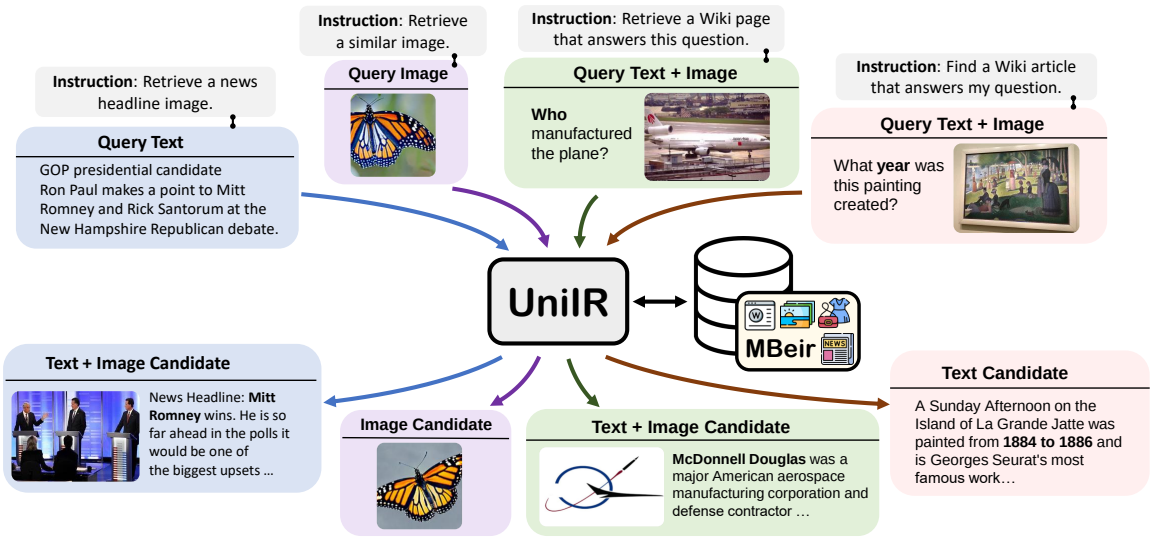


Figure 1. We build a universal multimodal information retriever UniIR through instruction tuning. UniIR is capable of accepting any form of query and instruction to retrieve information in any modality.

Abstract

Existing information retrieval (IR) models often assume a homogeneous format, limiting their applicability to diverse user needs, such as searching for images with text descriptions, searching for a news article with a headline image, or finding a similar photo with a query image. To approach such different information-seeking demands, we introduce UniIR, a unified instruction-guided multimodal retriever capable of handling eight distinct retrieval tasks across modalities. UniIR, a single retrieval system jointly trained on ten diverse multimodal-IR datasets, interprets user instructions to execute various retrieval tasks, demonstrating robust performance across existing datasets and zero-shot generalization to new tasks. Our experiments highlight that multi-task training and instruction tuning are keys to UniIR's generalization ability. Additionally, we construct the M-BEIR, a multimodal retrieval benchmark with comprehensive results, to standardize the evaluation of universal multimodal information retrieval.

1. Introduction

Information retrieval (IR) is a pivotal task that involves sourcing relevant information from vast data collections to meet specific user requirements [50]. This process has become increasingly important with the advent of generative AI [3, 9, 49, 58], as it not only enables precise attribution but also mitigates the risk of inaccuracies and fabrications in generated content [2, 48]. Despite the crucial role of IR in the current technological landscape, much of the existing literature—particularly within the realm of multimodal IR—remains narrow in scope, focusing mainly on homogeneous retrieval scenarios with pre-defined format, and oftentimes within a single domain. For example, MSCOCO [35] considers retrieving Flickr images via text caption, while EDIS [39] considers retrieving news headline images with news title. Such a homogeneous setting is insufficient to accommodate users' diverse information-seeking needs, which often transcends domains and modalities. For instance, while some users may search for web images through textual queries, others might use a photo

制造捏造

时常的

之前的多模态信息检索，需要使用特定领域数据集在CLIP等模型上进行微调

of a dress along with text input like "similar styles" or "color in red" to find similar fashion products for that specific dress. The current suite of multimodal retrieval systems falls short in its capacity to accommodate these diverse user demands, limited to task-specific fine-tuning of a pre-trained CLIP [47] model. In recognition of these limitations, a compelling need arises to conceptualize and develop a more flexible, general-purpose neural retriever that bridges different domains, modalities, and retrieval tasks to serve the diverse needs of users.

In this paper, we propose the UniIR framework to learn a single retriever to accomplish (possibly) any retrieval task. Unlike traditional IR systems, UniIR needs to follow the instructions to take a heterogeneous query to retrieve from a heterogeneous candidate pool with millions of candidates in diverse modalities. To train UniIR models, we construct M-BEIR, a benchmark of instruction following multimodal retrieval tasks building on existing 10 diverse datasets and unifying their queries and targets in a unified task formulation. The query instructions are curated to define the user's retrieval intention, thereby guiding the information retrieval process. We train different UniIR models based on pre-trained vision-language models like CLIP [47] and BLIP [33] on 300K training instances in M-BEIR with different multimodal fusion mechanisms (score-level fusion and feature-level fusion). We show that UniIR models are able to follow instructions precisely to retrieve desired targets from a heterogeneous candidate pool. Our best UniIR model is based on CLIP with score fusion, which not only achieves very competitive results on fine-tuned datasets but also generalizes to held-out datasets (Figure 6). Our ablation study reveals two insights: (1) Multi-task training in UniIR(BLIP) is beneficial, which leads to +9.7 improvement in terms of recall@5 over single-task training (Table 6); (2) Instruction tuning is critical to help models generalize to unseen retrieval datasets and leads to +10 improvement in terms of recall@5 (Figure 5).

Our contributions are summarized as follows:

- UniIR Framework: A universal multimodal information retrieval framework designed to integrate various multimodal retrieval tasks into a cohesive system.
- M-BEIR: A large-scale multimodal retrieval benchmark that assembles 10 diverse datasets from multiple domains, encompassing 8 distinct multimodal retrieval tasks.
- We introduce UniIR models, which are universal retrievers trained on M-BEIR, setting a foundational baseline for future research. Additionally, we evaluated the zero-shot performance of SOTA vision-language pre-trained models on the M-BEIR benchmark.

2. UniIR Framework

2.1. Problem Definition

In a universal multimodal search engine, users can initiate various search tasks based on their specific needs. These tasks involve different types of queries and retrieval candidates. The query $\mathbf{q}$ could be in text q_{t}, image q_{i} or even image-text pair $(q_{\text{i}}, q_{\text{t}})$, while the retrieval candidate $\mathbf{c}$ could also be text c_{t}, image c_{i} or an image-text pair $(c_{\text{i}}, c_{\text{t}})$. Eight existing retrieval tasks are being defined in Table 1. Please note that the compositional query, (q_i, q_t), typically involves a text-based question q_t about an image q_i. On the other hand, a compositional target, (c_i, c_t), usually includes an image c_i accompanied by a descriptive text c_t, providing contextual information.

To accommodate different retrieval intentions, we introduce a language task instruction $q_{\texttt{inst}}$ to represent the intention of the retrieval task. This instruction clearly defines what the search aims to find, whether seeking images, text, or a mix of both, and specifies the relevant domain. Further information can be found in Section 3. More formally, we aim to build a unified retriever model f capable of taking any type of query to retrieve any type of target specified by the instruction $q_{\texttt{inst}}$:

$$c^* = \arg\max_{\{\mathbf{c}\}\in\mathcal{C}} [f(\mathbf{q}, q_{\texttt{inst}})^T \cdot f(\mathbf{c})]$$

Here, $\mathcal{C}$ denotes the heterogeneous candidate pool, $f(\cdot)$ is the function we are optimizing for maximum dot-product retrieval, and c^* is the predicted result.

By including task instructions, we unify different multimodal retrieval tasks into a single framework, thus enabling us to build a general-purpose multimodal retriever. Furthermore, instruction fine-tuned language models have shown the capability to perform zero-shot generalization to unseen tasks by following instructions. However, applying this concept of zero-shot generalization to the multimodal retrieval domain faces challenges due to the lack of existing datasets tailored for this purpose. To address this gap, we are creating a comprehensive, unified dataset named M-BEIR, which is detailed in Section 3. M-BEIR will serve as a foundational resource for exploring and advancing the capabilities of multi-modal retrieval models.

2.2. UniIR Model

In this section, we present the UniIR model, our unified multimodal information retrieval system. The UniIR model is adept at handling distinct retrieval tasks simultaneously. We experimented with two multimodal fusion mechanisms for UniIR, namely score-level fusion [41] and feature-level fusion [19, 39]. To explore the effectiveness of these approaches, we adapted pre-trained models such as CLIP [47] and BLIP [33] for our purposes as follows.

Task (query → candidate)	Dataset	Instruction (shown 1 out of 4)	Domain	Train	Dev	Test	Pool
1. $q_t \rightarrow c_i$	VisualNews [37]	Identify news-related image match with the description	News	99K	20K	20K	542K
	MSCOCO [35]	Find an everyday image match with caption	Misc.	100K	24.8K	24.8K	5K
	Fashion200K [18]	Based on fashion description, retrieve matched image	Fashion	15K	1.7K	1.7K	201K
2. $q_t \rightarrow c_t$	WebQA [6]	Find an paragraph from Wikipedia to answer the question	Wiki	16K	1.7K	2.4K	544K
3. $q_t \rightarrow (c_i, c_t)$	EDIS [39]	Find a news image matching with the caption	News	26K	3.2K	3.2K	1M
	WebQA [6]	Find a Wiki image that answer the question	Wiki	17K	1.7K	2.5K	403K
4. $q_i \rightarrow c_t$	VisualNews [37]	Provide a news-related caption for the displayed image	News	100K	20K	20K	537K
	MSCOCO [35]	Find a caption describe the an image	Misc.	113K	5K	5K	25K
	Fashion200K [18]	Find a description for the fashion item in the image	Fashion	15K	4.8K	4.8K	61K
5. $q_i \rightarrow c_i$	NIGHTS [15]	Find an image that is identical to the given image	Misc.	16K	2K	2K	40K
6. $(q_i, q_t) \rightarrow c_t$	OVEN [19]	Retrieve a Wiki text that answer the given query about the image	Wiki	150K	50K	50K	676K
	InfoSeek [10]	Find an article that answers the given question about the image	Wiki	141K	11K	11K	611K
7. $(q_i, q_t) \rightarrow c_i$	FashionIQ [56]	Find an image to match the fashion image and style note	Fashion	16K	2K	6K	74K
	CIRR [40]	I'm looking for a similar everyday image with the described changes	Misc.	26K	2K	4K	21K
8. $(q_i, q_t) \rightarrow (c_i, c_t)$	OVEN [19]	Find a Wiki image-text pair to answer a question regarding an image	Wiki	157K	14.7K	14.7K	335K
	InfoSeek [10]	Find a Wiki image-text pair to answers my question about this image	Wiki	143K	17.6K	17.6K	481K
	10 datasets	64 instructions	4 domains	1.1M	182K	190K	5.6M

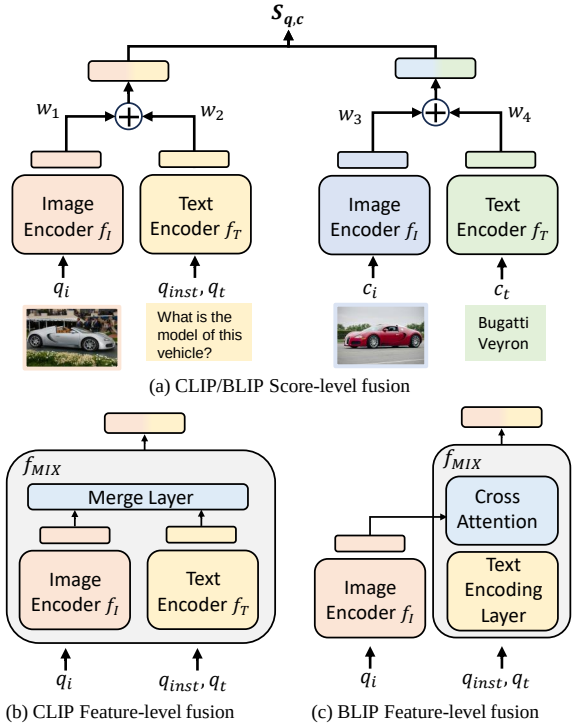
Table 1. The overview of M-BEIR training/validation/test set. More detailed query instruction design can be found in Appendix.

Score-level Fusion. As illustrated in Figure 2(a), the score-level fusion variants for CLIP and BLIP (denoted as CLIP_{SF} and BLIP_{SF}) employ distinct encoders for vision and text. Specifically, the vision encoder is marked as f_i and the uni-modal text encoder as f_t. In these methods, both image and text inputs (whether from a query or a target) are processed into two individual vectors. These vectors undergo a weighted sum to form a unified representation vector. This process is mathematically represented as $f(q_\text{i}, q_\text{t}, q_\texttt{inst}) = w_1 f_I(q_\text{i}) + w_2 f_T(q_\text{t}, q_\texttt{inst})$ for queries and $f(c_\text{i}, c_\text{t}) = w_3 f_I(c_\text{i}) + w_4 f_T(c_\text{t})$ for targets. Therefore, the similarity score between a query $\mathbf{q}$ and a target $\mathbf{c}$ is calculated as a weighted sum of the within-modality and cross-modality similarity scores:

$$
\begin{aligned}
s_{\mathbf{q},\mathbf{c}} &= f(q_\text{i}, q_\text{t}, q_\texttt{inst})^T \cdot f(c_\text{i}, c_\text{t}) \\
&= w_1 w_3 f_I(q_\text{i})^T f_I(c_\text{i}) + w_2 w_4 f_T(q_\text{t}, q_\texttt{inst})^T f_T(c_\text{t}) \\
&+ w_1 w_4 f_I(q_\text{i})^T f_T(c_\text{t}) + w_2 w_3 f_T(q_\text{t}, q_\texttt{inst})^T f_I(c_\text{i})
\end{aligned}
$$

w_1, w_2, w_3, w_4 is a set of learnable parameters that reflects importance weights.

Feature-level Fusion. Contrasting the approach of processing uni-modal data separately, feature-level fusion integrates features during the encoding phase. This fusion method computes a unified feature vector for multi-modal queries or candidates using mixed-modality attention layers. As illustrated in Figure 2 (b), for the CLIP feature-level fusion (CLIP_{FF}), we have enhanced the pre-trained vision encoder f_I and text encoder f_T with a 2-layer Multi-Modal Transformer, which follows the same architecture as T5



(a) CLIP/BLIP Score-level fusion

(b) CLIP Feature-level fusion

(c) BLIP Feature-level fusion

Figure 2. (a) Score-level fusion encodes each modality into a single feature; (b) CLIP feature-level fusion (CLIP_{FF}) fuses two modalities into a single feature with a mix-modality transformer layer; (c) BLIP feature-level fusion (BLIP_{FF}) adopts cross-attention to output a single feature vector.

Transformer, forming a mixed-modality encoder f_{MIX}. In the case of BLIP feature-level fusion (BLIP_{FF}), the process begins with the extraction of image embeddings through the

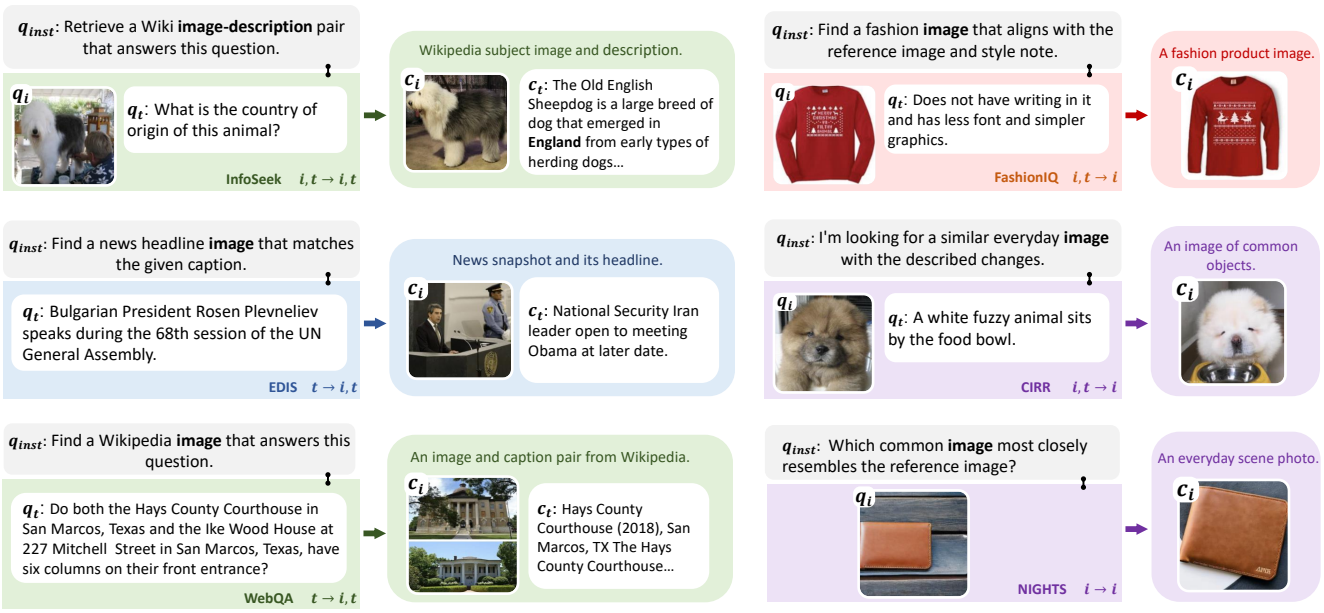


Figure 3. Examples of six query instances in the M-BEIR dataset. Each example query instance includes a query $\mathbf{q}$, a human-annotated natural language instruction q_{inst}, and a positive(relevant) candidate $\mathbf{c}^+$.

vision encoder f_I. These embeddings are then integrated with text embeddings through the cross-attention layers of BLIP's image-grounded text encoder, also labeled as f_{MIX}. In both CLIP_{FF} and BLIP_{FF}, the output from f_{MIX} is a comprehensive feature vector that combines information from both image and text modalities. The final representations for the query and target, denoted as $f_{\text{MIX}}(q_{\text{i}}, q_{\text{t}}, q_{\texttt{inst}})$ and $f_{\text{MIX}}(c_{\text{i}}, c_{\text{t}})$ respectively, are obtained separately but using the same f_{MIX}. The similarity score between the query and the target is then calculated by:

$$s_{\mathbf{q},\mathbf{c}} = f_{\text{MIX}}(q_{\text{i}}, q_{\text{t}}, q_{\texttt{inst}})^T \cdot f_{\text{MIX}}(c_{\text{i}}, c_{\text{t}})$$

We fine-tuned the above-detailed four types of model variants on the M-BEIR training data (detail in Section 3), employing the query-target contrastive objective. To adhere to a uniform instruction tuning format, instructions $q_{\texttt{inst}}$ were integrated as prefixes to the text query q_{t}. See examples in Figure 3. We input padding tokens for queries or candidates missing either image or text modalities.

3. M-BEIR Benchmark

To train and evaluate unified multimodal retrieval models, we build a large-scale retrieval benchmark named M-BEIR (**M**ultimodal **BE**nchmark for **I**nstructed **R**etrieval). The M-BEIR benchmark comprises eight multimodal retrieval tasks and ten datasets from a variety of domains and image sources. Each task is accompanied by human-authored instructions, encompassing 1.5 million queries and a pool of 5.6 million retrieval candidates in total (see Table 1).

3.1. Data Format

To unify multimodal retrieval tasks, which consist of different modalities in the source query and target candidate, each task in M-BEIR includes queries $\mathcal{Q} = \{\mathbf{q_1}, \mathbf{q_2}, ...\}$, a set of candidates $\mathcal{C} = \{\mathbf{c_1}, \mathbf{c_2}, ...\}$, where $\mathbf{q}$ and $\mathbf{c}$ both support text and image modality, and a human-authored instruction $q_{\texttt{inst}}$ is provided to specify the intent of the retrieval task. Each query instance in the M-BEIR dataset includes a query $\mathbf{q}$, an instruction $q_{\texttt{inst}}$, a list of relevant(positive) candidate data $\mathbf{c}^+$ and a list of potentially available irrelevant(negative) candidate data $\mathbf{c}^-$. See examples in Figure 3. Every M-BEIR query instance has at least one positive candidate data and possibly no negative candidate data. Our default retrieval setting is that the model needs to retrieve the positive candidates from a heterogeneous pool of candidates in all different modalities and domains.

3.2. Dataset Collection

The M-BEIR benchmark encompasses various domains: everyday imagery, fashion items, Wikipedia entries, and news articles. It integrates 8 multimodal retrieval tasks by leveraging a variety of datasets.

Data Selection. To build a unified instruction-tuned multimodal retrieval model and comprehensive evaluation benchmark, we aim to cover a wide range of multimodal tasks, domains, and datasets. These include retrieval-focused datasets (OVEN [19], EDIS [39], CIRR [40] and FashionIQ [56]), image-caption datasets (MS-COCO [35], Fashion200K [18], VisualNews [37]), image-similarity measurement dataset (NIGHTS [15]), along with retrieval-

Task	Dataset	Zero-shot	Multi-task (✗ instruction)					UniIR(✓ instruction)				
		BLIP2	CLIP_{SF}	CLIP_{FF}	BLIP_{SF}	BLIP_{FF}	$\text{BLIP}_{\text{FF, 384}}$	CLIP_{SF} (Δ)	CLIP_{FF} (Δ)	BLIP_{SF} (Δ)	BLIP_{FF} (Δ)	$\text{BLIP}_{\text{FF, 384}}$ (Δ)
1. $q_t \to c_i$	VisualNews	0.0	12.7	8.8	5.0	8.3	10.5	**42.6** (+29.9)	28.8 (+20.0)	20.9 (+15.8)	23.0 (+14.8)	26.5 (+16.0)
	MSCOCO	0.0	27.3	24.6	22.9	27.7	35.2	**77.9** (+50.6)	74.7 (+50.1)	71.6 (+48.7)	75.6 (+47.8)	75.7 (+40.4)
	Fashion200K	0.0	5.9	5.9	5.7	9.0	13.1	17.8 (+11.9)	15.5 (+9.7)	24.3 (+18.6)	25.4 (+16.4)	**26.7** (+13.6)
2. $q_t \to c_t$	WebQA	35.2	82.3	67.9	74.4	76.1	76.9	**84.7** (+2.5)	78.4 (+10.6)	78.9 (+4.4)	79.5 (+3.4)	79.2 (+2.4)
3. $q_t \to (c_i, c_t)$	EDIS	0.0	41.1	38.3	33.6	36.0	38.5	**59.4** (+18.3)	50.0 (+11.7)	47.2 (+13.6)	50.3 (+14.4)	51.4 (+12.9)
	WebQA	0.0	68.2	62.5	73.2	74.7	75.2	78.8 (+10.6)	75.3 (+12.8)	76.8 (+3.6)	**79.7** (+5.0)	79.4 (+4.2)
4. $q_i \to c_t$	VisualNews	0.0	12.1	8.2	4.8	4.9	6.0	**42.8** (+30.7)	28.6 (+20.4)	19.4 (+14.6)	21.1 (+16.3)	22.9 (+16.9)
	MSCOCO	0.0	84.6	80.8	74.9	76.9	81.4	**92.3** (+7.8)	89.0 (+8.2)	88.2 (+13.4)	88.8 (+11.9)	90.1 (+8.7)
	Fashion200K	0.0	1.2	1.3	2.6	3.6	4.0	17.9 (+16.7)	13.7 (+12.4)	24.3 (+21.7)	27.6 (+24.1)	**28.4** (+24.4)
5. $q_i \to c_i$	NIGHTS	24.0	31.0	30.8	32.9	31.3	32.5	32.0 (+1.0)	31.9 (+1.2)	33.4 (+0.4)	33.0 (+1.6)	**33.7** (+1.3)
6. $(q_i, q_t) \to c_t$	OVEN	0.0	36.8	31.6	33.2	37.7	39.2	39.2 (+2.4)	34.7 (+3.1)	35.2 (+2.0)	38.7 (+1.0)	**40.7** (+1.5)
	InfoSeek	0.0	18.3	15.4	11.9	17.8	17.1	**24.0** (+5.8)	17.5 (+2.1)	16.7 (+4.8)	19.7 (+1.9)	19.2 (+2.0)
7. $(q_i, q_t) \to c_i$	FashionIQ	3.9	22.8	19.7	26.1	28.1	29.0	24.3 (+1.5)	20.5 (+0.9)	26.2 (+0.1)	28.5 (+0.5)	**29.8** (+0.9)
	CIRR	6.2	32.0	32.7	36.7	45.1	47.4	43.9 (+11.9)	40.9 (+8.2)	43.0 (+6.3)	**51.4** (+6.3)	51.1 (+3.8)
8. $(q_i, q_t) \to (c_i, c_t)$	OVEN	13.8	58.7	50.1	51.0	51.6	53.1	**60.2** (+1.5)	55.8 (+5.7)	51.8 (+0.8)	57.8 (+6.2)	59.5 (+6.4)
	InfoSeek	11.4	42.3	31.5	23.0	25.4	25.2	**44.6** (+2.4)	36.8 (+5.3)	25.4 (+2.5)	27.7 (+2.3)	31.1 (+5.9)
	Average	5.9	36.1	31.9	32.0	34.6	36.5	**48.9** (+12.8)	43.3 (+11.4)	42.7 (+10.7)	45.5 (+10.9)	46.6 (+10.1)

Table 2. Benchmarking universal information retrieval on M-BEIR (5.6M candidates) with Recall@5 (except using Recall@10 for Fashion200K, FashionIQ). Δ: UniIR - Multi-task. 384 refers to image resolution. **Bold**: top-1 performance. Underline: top-2.

based VQA datasets (InfoSeek [10], WebQA [6]). These datasets, originally designed for different purposes, are effectively repurposed as retrieval tasks within the M-BEIR benchmark. In the case of image-caption datasets, we repurpose the image-caption pair as the retrieval task following MS-COCO. For the other datasets, we adopt original queries and use the annotated gold candidates as positive candidates $\mathbf{c}^+$ and annotated hard negatives as irrelevant candidates $\mathbf{c}^-$. We also adopt the provided candidate pool. In total, M-BEIR covers 8 different multimodal retrieval tasks and 4 domains with a global pool of 5.6 million candidates. See Table 3 for the full dataset list. To ensure data balance in our benchmark, we trim down candidate pools and instances from the larger datasets such as VisualNews, OVEN, and InfoSeek, which originally contained 1 to 6 million instances, significantly larger than other datasets. To facilitate training, validation, and testing, we use the original dataset splits from each dataset. If the dataset only releases a validation set, we hold out a part of the training data to use for validation and report results on the original validation set. Otherwise, we report results using the test set. More details on data processing can be found in the Appendix.

Instruction Annotation Guideline. One of the key components of the success of instruction-tuning is the diverse instructions that specify the intention of the task [12, 55]. To design instructions for multimodal retrieval tasks, we took inspiration from the instruction schema in TART [1]. Our M-BEIR instruction describes a multimodal retrieval task by intent, domain, query modality, and target candidate modality. Specifically, intent describes how the retrieved resources are related to the query. The domain defines the expected resource of the target candidate, such as Wikipedia or fashion products. For a text-to-image retrieval dataset like Fashion200K [18], our instruction would be: "Based on the following fashion description, retrieve the best matching image." More examples in Table 1 and Figure 3. Following the instruction annotation guideline, we authored 4 instructions for each query in every retrieval task. The full list of instructions is in the Appendix Table 9 and Table 10.

3.3. Evaluation Metrics

We follow the standard retrieval evaluation metric, recall@k, used for MSCOCO and report results for all datasets. Specifically, we adhere to the recall implementation of CLIP [47]/BLIP [33] for MSCOCO, which counts the retrieved instance as correct if it overlaps with relevant instances. We mainly report Recall@5 for all datasets except Fashion200K and FashionIQ, following the prior work [56] to report Recall@10. Full results of Recall@1/5/10 can be found in the Appendix Table 11.

4. Experiments

In our experiments, we assess a variety of multimodal retrieval models on the M-BEIR dataset, leveraging pre-trained vision-language transformer models. We use publicly available checkpoints, as listed in Table 2. Our evaluation encompasses both SoTA models, fine-tuned baselines and UniIR models(detailed in Section 2) under two retrieval scenarios: (1) retrieving from the M-BEIR 5.6 million candidate pool, which consists of the retrieval corpus from all tasks, and (2) retrieving from a task-specific pool (with homogeneous candidates) provided by the original dataset, which is to enable fair comparison with existing SoTA retrievers. We name this task-specific pool as $\text{M-BEIR}_{\text{local}}$. The retrieval process involves a two-step pipeline. Firstly, we extract multimodal feature vectors for all the queries and

Dataset	Domain	Task	Query Instruction $q_{\texttt{inst}}$	Query Image q_i	Query Text q_t
EDIS	News	3. $q_t \rightarrow (c_i, c_t)$	Find a news headline image that matches the provided caption.	-	Barack Obama with Germany's chancellor Angela Merkel at the Brandenburg Gate Berlin on 19 June.

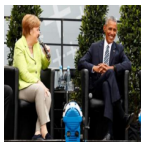
Model	Rank 1 (c_i, c_t)	Rank 2 (c_i, c_t)	Rank 3 (c_i, c_t)	Rank 4 (c_i, c_t)	Rank 5 (c_i, c_t)
UniIR (CLIP_{SF}) ✓inst	✗  Obama Speaks at a Berlin Event With Angela Merkel.	✓  Obama's Berlin visit to coincide with Trump in Brussels - Barack Obama.	✗  Obama meets Germany's Merkel at chancellery in Berlin.	✓  President Obama Speaks to the People of Berlin from the Brandenburg Gate.	✓  When Barack Obama visited Berlin two years ago, he charmed a city.
Multi-task (CLIP_{SF}) ✗inst	✗ President Obama stands next to German Chancellor Angela Merkel in front of Brandenburg Gate in Berlin on June 19.	✗ President Obama and German Chancellor Angela Merkel in 2011.	✓  When Barack Obama visited Berlin two years ago, he charmed a city.	✗ Barack Obama with German Chancellor Angela Merkel at the G20 summit in November.	✗ US president Barack Obama at the Brandenburg Gate.
Zero-shot (BLIP2) ✗inst	✗ President Obama stands next to German Chancellor Angela Merkel in front of Brandenburg Gate in Berlin on June 19.	✗ US President Barack Obama waves next to German Chancellor Angela Merkel before they deliver speeches to invited guests in front of the Brandenburg Gate at Pariser Platz in Berlin on June 19 2013 during the official visit of the US President Barack Obama walks in ...	✗ Barack Obama with German chancellor Angela Merkel at the G20 summit in November.	✗ BERLIN GERMANY JUNE 19 US President Barack Obama meets German Chancellor Angela Merkel for bilateral talks at the Chancellery on June 19 2013 in Berlin Germany Obama is visiting Berlin for the first time during his presidency and his speech at the Brandenburg Gate...	✗ President Obama walks with Germany s Chancellor Angela Merkel in St Petersburg Russia on Sept 6.

Figure 4. Visualization of top 5 retrieved candidates from M-BEIR with 3 models on EDIS. Without instructions, zero-shot and multi-task training models mostly retrieve the wrong modality (text-only). UniIR retrieves candidates accurately with the right modality (image, text).

candidates in the pool. We then utilize FAISS [26], a powerful library for efficient similarity searches in dense vector spaces, to index and retrieve candidates.

4.1. Baselines

Zero-shot SoTA Retriever. We utilize pre-trained vision-language models such as CLIP (L-14)[47], SigLIP (L)[59], BLIP (L) [33], and BLIP2 [34] as our baseline feature extractors. The caveat is that these models cannot understand the intent of the retrieval task as the input is only query **q**, thus, they are expected to achieve low performance in the standard setting (1) with a heterogeneous candidate pool. We do not use q instructions as we found this even degrades zero-shot retrieval performance.

Single/Multi-task Fine-tuned Baselines. We fine-tune CLIP and BLIP on this specific dataset as our single-task baseline retrievers. We also fine-tune CLIP and BLIP jointly on all M-BEIR training data without incorporating instructions as our multi-task baseline retrievers. The model only takes in **q** and **c** using the query-target contrastive training objective to maximize the positive pair similarity while minimizing negative pair similarity.

Implementation Details. For all the CLIP and BLIP variants, we employ the largest checkpoint, i.e., ViT-L14 [14]. The default image resolution is 224×224 unless specified otherwise. We use a batch size of 105 for CLIP variants and 115 for BLIP variants. We adopt other hyperparameters as reported in the original implementations. For score fusion methods, we set $w_1 = w_2 = w_3 = w_4 = 1$ by default. All our experiments are conducted on a single node with 8 H100 GPUs. Further details can be found in the Appendix.

4.2. Experimental Results

We report the main results on M-BEIR in Table 2, where models retrieve candidates from the 5.6 M pool. We show that zero-shot models struggle to retrieve queried information from such a heterogeneous pool. We demonstrate instruction-tuning as a crucial component in Table 3 and show multimodal fusion architecture design insights in Table 5. Furthermore, we also conduct experiments to understand the zero-shot generalization of UniIR, where we would train UniIR on a subset of datasets and evaluate it on the held-out test set, which makes UniIR fairly comparable with other zero-shot retrievers.

Zero-shot retrievers cannot comprehend retrieval intention. We first benchmark four open-sourced cross-modal embedding models and found the recall values on most tasks are near zero. To demonstrate this, we have provided an example of BLIP2 in Table 2. These pre-trained models struggle to comprehend the task intention without the

警告，附加说明

目标函数，和CLIP预先训练的目标函数应该类似

guidance of instruction. For example, in the text-to-image retrieval task on MSCOCO, all zero-shot models retrieve text instances from the global pool, leading to 0% recall rate. This outcome is expected, given that similarity scores tend to be higher when the query and candidate come from the same modality. Furthermore, we observed that zero-shot models, for example, BLIP2 in Table 2 cannot effectively fuse modalities as the recall of WebQA drops from 35.2% to 0% when the retrieval candidates are image-text pairs instead of text snippets. In Figure 4, we present examples where BLIP2 retrieves distracting candidates from the wrong modality for an EDIS query.

	ZS	Multi.	UniIR	ZS	Multi.	UniIR
Task	CLIP	CLIP_{SF}	CLIP_{SF} (Δ)	BLIP	BLIP_{FF}	BLIP_{FF} (Δ)
1.	0.0	15.3	46.1 (+30.8)	0.0	15.0	41.3 (+26.3)
2.	32.1	82.3	84.7 (+2.5)	38.1	76.1	79.5 (+3.4)
3.	6.1	54.6	69.1 (+14.5)	0.0	62.0	65.0 (+3.0)
4.	0.0	32.6	51.0 (+18.4)	0.0	28.4	45.9 (+17.4)
5.	25.3	31.0	32.0 (+1.0)	25.1	31.3	33.0 (+1.6)
6.	0.0	27.5	31.6 (+4.1)	0.0	27.8	29.2 (+1.5)
7.	4.9	27.4	34.1 (+6.7)	4.8	36.6	40.0 (+3.4)
8.	23.3	50.5	52.4 (+1.9)	9.0	38.5	42.7 (+4.2)
Avg.	7.9	36.1	48.9 (+12.8)	5.7	34.6	45.5 (+10.9)

Table 3. **Experiments of instruction-tuning**. Retrieve from the M-BEIR (Recall@5). Δ: UniIR - Multi-task (Multi.).

Instruction-tuning improves retrieval on M-BEIR. To understand the benefit of instruction-tuning in UniIR, we present a comparison of UniIR with multi-task fine-tuned baselines in Table 3. Despite having the same architecture, UniIR models show significant improvement over baselines on M-BEIR. The average Recall@5 has increased by 12.8 and 10.9, respectively. We also discovered that the largest improvement was observed in cross-modality retrieval tasks 1 and 3. Without instructions, the multi-task baselines struggle to understand the task intention and tend to retrieve candidates from the same modality as the query. However, Instruction-tuning does not significantly improve within-modality retrieval tasks like 2 and 5 as these do not require the embedding model to understand intent.

UniIR can precisely follow instructions. To further demonstrate the advantages of UniIR over Multi-task fine-tuning baselines, we conducted an analysis of the retrieval error. The errors were classified into three categories: incorrect modality, incorrect domain, and other errors. The results are presented in Table 4. The Multi-task models showed a high error rate of 58.8% and 50.9% in retrieving instances with the wrong modality from the global pool. However, with instruction finetuning, UniIR models were able to successfully learn to retrieve intended modalities, resulting in a significant drop in error rate to 2.7% and 15.2%. In Figure 4, we show examples of incorrect modality errors by visualizing the top 5 retrieved candidates using zero-shot, multi-task and UniIR models on one of EDIS queries. Specifically, the zero-shot model (BLIP2) and multi-task model (CLIP_{SF}) mostly retrieve distracting candidates from the wrong modality (c_t), while UniIR (CLIP_{SF}) retrieves all positive candidates from the right modality (c_i, c_t). More examples can be found in Appendix Figure 7-16.

	Multi-task		UniIR	
Error Types	CLIP_{SF}	BLIP_{FF}	CLIP_{SF}	BLIP_{FF}
✗ modality	58.8%	50.9%	2.7%	15.2%
✗ domain	0.3%	0.5%	0.1%	0.0%
Other	40.9%	48.6%	97.2%	84.8%

Table 4. Error analysis on M-BEIR.

	Zero-shot		UniIR			
Task	CLIP	BLIP	CLIP_{FF}	CLIP_{SF} (Δ)	BLIP_{SF}	BLIP_{FF} (Δ)
1.	0.0	0.0	39.7	46.1 (+6.4)	38.9	41.3 (+2.4)
2.	32.1	38.1	78.4	84.7 (+6.3)	78.9	79.5 (+0.7)
3.	6.1	0.0	62.7	69.1 (+6.4)	62.0	65.0 (+3.0)
4.	0.0	0.0	43.8	51.0 (+7.3)	44.0	45.9 (+1.9)
5.	25.3	25.1	31.9	32.0 (+0.1)	33.4	33.0 (–0.4)
6.	0.0	0.0	26.1	31.6 (+5.5)	26.0	29.2 (+3.2)
7.	4.9	4.8	30.7	34.1 (+3.4)	34.6	40.0 (+5.4)
8.	23.3	9.0	46.3	52.4 (+6.1)	38.6	42.7 (+4.1)
Avg	7.9	5.7	43.3	48.9 (+5.7)	42.7	45.5 (+2.8)

Table 5. **Experiments of multimodal feature fusion architecture design**. Retrieve from the M-BEIR (Recall@5). Δ: difference between two model architectures.

UniIR can generalize to unseen retrieval tasks. During the multi-task fine-tuning stage of UniIR, we excluded three datasets (WebQA, OVEN, CIRR) and fine-tuned UniIR models and multi-task baselines on the remaining M-BEIR datasets. At test time on the M-BEIR global pool, we evaluated the zero-shot performance of all fine-tuned models, as well as SoTA pre-trained retrievers (CLIP and BLIP) on the three held-out datasets. In Figure 5, we compared the average performance of SoTA (CLIP and BLIP) retrievers, the average performance of multi-task fine-tuned baselines Multi-task(CLIP_{SF}) and Multi-task(BLIP_{FF}), and the average performance of UniIR (CLIP_{SF}) and UniIR (BLIP_{FF}). Our results indicate two main findings. Firstly, UniIR models outperform SoTA retriever baselines by a significant margin on held-out datasets during zero-shot evaluation. Secondly, we demonstrate that UniIR models, which incorporate instruction-tuning, exhibit superior generalization abilities on unseen tasks and datasets compared to their multi-task counterparts without instructions.

Aligning the model architecture with pre-training. In Table 5, we compare fusion architecture designs between score-fusion and feature-fusion, where score-fusion is native to the CLIP model (i.e., CLIP_{SF}) and feature-fusion

		SoTA Zero-Shot				ST	MT	UniIR	ST	MT	UniIR
Task	**Dataset**	CLIP	SigLIP	BLIP	BLIP2	CLIP_{SF}	CLIP_{SF}	CLIP_{SF} (Δ_s)	BLIP_{FF}	BLIP_{FF}	BLIP_{FF} (Δ_s)
1. $q_t \rightarrow c_i$	VisualNews	43.3	30.1	16.4	16.7	43.5	40.6	42.6 (–0.9)	20.0	22.8	23.4 (+3.4)
	MSCOCO	61.1	75.7	74.4	63.8	80.4	79.9	81.1 (+0.7)	77.3	78.3	79.7 (+2.3)
	Fashion200K	6.6	36.5	15.9	14.0	10.7	16.8	18.0 (+7.4)	17.1	25.8	26.1 (+9.0)
2. $q_t \rightarrow c_t$	WebQA	36.2	39.8	44.9	38.6	81.7	83.7	84.7 (+3.1)	67.5	77.9	80.0 (+12.5)
3. $q_t \rightarrow (c_i, c_t)$	EDIS	43.3	27.0	26.8	26.9	58.8	57.4	59.4 (+0.6)	38.2	51.2	50.9 (+12.7)
	WebQA	45.1	43.5	20.3	24.5	76.3	76.7	78.7 (+2.5)	67.8	79.2	79.8 (+11.9)
4. $q_i \rightarrow c_t$	VisualNews	41.3	30.8	17.2	15.0	42.7	40.0	43.1 (+0.4)	22.4	20.9	22.8 (+0.3)
	MSCOCO	79.0	88.2	83.2	80.0	89.8	90.3	92.3 (+2.6)	86.0	85.8	89.9 (+3.9)
	Fashion200K	7.7	34.2	19.9	14.2	12.0	18.4	18.3 (+6.3)	15.6	27.4	28.9 (+13.3)
5. $q_i \rightarrow c_t$	NIGHTS	26.1	28.9	27.4	25.4	33.5	31.1	32.0 (–1.5)	30.4	31.5	33.0 (+2.6)
6. $(q_i, q_t) \rightarrow c_t$	OVEN	24.2	29.7	16.1	12.2	45.4	46.6	45.5 (+0.1)	33.8	42.8	41.0 (+7.2)
	InfoSeek	20.5	25.1	10.2	5.5	23.5	28.3	27.9 (+4.4)	18.5	23.9	22.4 (+3.9)
7. $(q_i, q_t) \rightarrow c_i$	FashionIQ	7.0	14.4	2.3	4.4	25.9	23.2	24.4 (–1.5)	3.0	28.4	29.2 (+26.2)
	CIRR	13.2	22.7	10.6	11.8	52.0	38.7	44.6 (–7.3)	13.9	48.6	52.2 (+38.2)
8. $(q_i, q_t) \rightarrow (c_i, c_t)$	OVEN	38.8	41.7	27.4	27.3	66.2	69.0	67.6 (+1.4)	49.9	56.3	55.8 (+5.9)
	InfoSeek	26.4	27.4	16.6	15.8	47.4	49.2	48.9 (+1.5)	32.3	32.9	33.0 (+0.7)
-	Average	32.5	37.2	26.8	24.8	49.4	49.4	50.6 (+1.2)	37.1	45.8	46.8 (+9.7)

Table 6. Multi-task (MT) instruction-tuning experiments on $\text{M-BEIR}_{\text{local}}$. We report Recall@5 results of zero-shot retrieval, single-task (ST) fine-tuning, and UniIR (with or without instructions) on $\text{M-BEIR}_{\text{local}}$ except for Fashion200K and FashionIQ where we report Recall@10. Retrieval is conducted from $\text{M-BEIR}_{\text{local}}$ (single) candidate pools. Δ_s: absolute difference to single task fine-tuning.

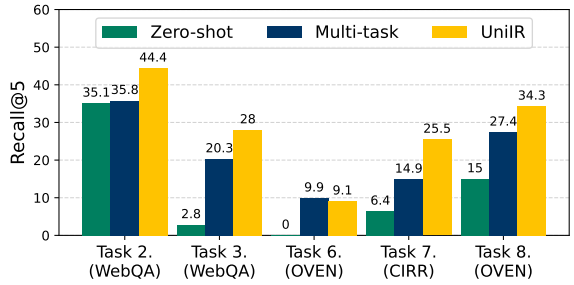


Figure 5. **Held-out dataset generalization experiments** on M-BEIR: we train a Multi-task and a UniIR model on 7 held-in datasets and test on 3 held-out datasets (WebQA, OVEN, CIRR) from the M-BEIR. Results are averaged over CLIP_{SF} and BLIP_{FF}.

is native to BLIP with its pre-trained cross-attention transformer encoder (i.e., BLIP_{FF}). By adhering to the pre-training architecture design, we show that the fine-tuned UniIR models attain higher Recall@5 scores for each task, with an average improvement of 5.7 and 2.8 for CLIP and BLIP, respectively. Furthermore, avoid adding randomly initialized layers during fine-tuning, such as the T5 mix-modality layers used in CLIP_{FF}. These layers are not pre-trained and can lead to overfitting - specifically for tasks like OVEN/InfoSeek (task 8). The comparison of CLIP_{FF} and CLIP_{SF} on task 8 reveals a significant drop in performance, from 52.4 to 46.3, for the latter.

4.3. Comparison with Existing Methods

To compare UniIR with existing retrievers, we also evaluate the homogeneous setting where the retriever only needs to retrieve from the task-specific pool, which is more consistent with the traditional IR setup. Additionally, we conducted held-out experiments to examine UniIR's zero-shot generalization ability on task-specific pools $\text{M-BEIR}_{\text{local}}$ in comparison to baseline models.

UniIR vs Zero-shot Retrievers. In Table 6, we demonstrate that while SigLIP attains the highest average value of zero-shot SoTA retrievers with an average value of 37.2% on R@5, our UniIR models (CLIP_{SF}) and (BLIP_{FF}) surpass it by a significant margin, with average R@5 values of 50.6% and 46.8% respectively.

UniIR vs Single-task Tuning. Table 6 demonstrates the advantages of multi-task instruction-tuning in the UniIR framework over single-task fine-tuning. Our findings indicate that UniIR (BLIP_{FF}) greatly outperforms its single-task counterpart by an average of 9.7% on R@5, and exhibits significant improvements on task 7 compositional image retrieval such as CIRR with 48.6% compared to 13.9%. UniIR (CLIP_{SF}) also demonstrates an overall improvement of 1.2%, particularly on Fashion200K and Infoseek. In contrast, we observed that the multi-task training without instructions would not lead to such improvements on average for CLIP_{SF}, as it remained at 49.4%.

Generalization Performance on Held-Out Datasets and Tasks In Figure 6, we showed the average zero-shot performance of SoTA CLIP and BLIP retrievers, the average zero-shot performance of multi-task fine-tuned baselines, and the average zero-shot performance of UniIR (CLIP_{SF}) and UniIR (BLIP_{FF}) on 3 held-out datasets on $\text{M-BEIR}_{\text{local}}$. The

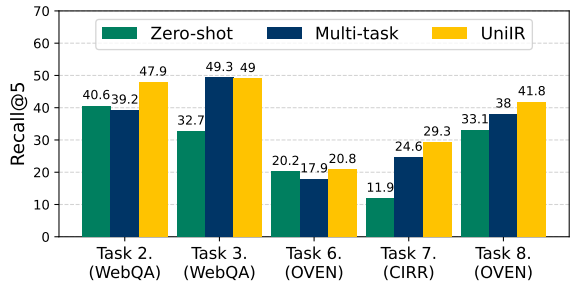


Figure 6. **Held-out dataset generalization experiments** on M-BEIR$_{\text{local}}$: we train a Multi-task and a UniIR model on 7 held-in datasets and test on 3 held-out datasets (WebQA, OVEN, CIRR). Results are averaged over CLIP$_{\text{SF}}$ and BLIP$_{\text{FF}}$.

UniIR models exhibit superior generalization ability on unseen tasks and datasets. As shown in Figure 6, UniIR models consistently outperform the SoTA retrievers and multi-task training baselines over 3 held-out datasets across 5 tasks. On the other hand, Multi-task training without using instruction shows moderate improvements over the SoTA retriever baselines and performs even worse in tasks such as WebQA (task 2) and OVEN (task 6).

5. Related Work

Multimodal Information Retrieval. In recent years, the field of cross-modal information retrieval has seen significant exploration, with a particular emphasis on image-to-text matching. Datasets such as MSCOCO [27] and Flickr30k [46] have become standard benchmarks for evaluating the progress of pre-trained vision-language models such as ALIGN [25], VILT [28], ALBEF [32], MURAL [24], and ImageBind [16]. However, fine-grained image retrieval often hinges on the ability to articulate intents through text, presenting challenges in multimodal queries [7] such as ReMuQ [42]. While the text-to-text retrieval benchmark BEIR [52] has advanced research in building generalized zero-shot text retrieval systems, a unified multimodal information retrieval benchmark covering a diverse range of tasks remains absent. We hope that the introduction of the M-BEIR will accelerate progress toward more general multimodal information retrieval models.

Retrieval-augmented Models. Retrieval augmentation has been studied extensively in the past few years. ORQA [29], RAELM [17], RAG [30] and FID [22] are among the earliest work to learn retriever and language model jointly from weakly supervised dataset. Later on, RETRO [4] scaled the retrieval-augmented training to large-scale language models to show great performance gain with a relatively small-sized language model. ATLAS [23] further extends the idea of retrieval-augmented training to few-shot learning and shows performance gain on broader knowledge-intensive tasks. These works are mostly focused on retrieving text paragraphs to augment language models. Later on, RA-C3M[58], REVEAL [20], and MuRAG [8] have shown the advantage of retrieving multimodal content from Wikipedia to answer visually information-seeking questions. The multimodal augmentation idea is also applied to image generation with several works like KNN-difussion[49] Re-imagen [9], and RA-diffusion[3]. The closest to our work is the TART [1], which also aims to build a single retriever to retrieve different content based on the instruction. However, TART is still only focused on text-to-text modality.

Instruction tuning. Instruction-tuned models, where models are trained to follow user instructions, have emerged as a significant area of research in large language models (LLMs) [11, 44, 53]. FLAN and FLAN-T5 [12, 55] have demonstrated capabilities to generalize to unseen natural language tasks [43] and InstructGPT [45] further illustrates how instruction tuning can align language models more closely with users' intentions. Recently, visual instruction tuning has been explored in vision-language tasks such as visual question answering with models such as InstructBLIP [13] and LLaVA [38] or datasets such as the MultiInstruct [57]. On image diffusion models, InstructPix2Pix [5] and MagicBrush [60] dataset show how the diffusion model can follow instructions to edit images. However, the most closely related retrieval-augmented models, such as InstructRETRO [54], RA-DIT [36], as well as embedding models like OneEmbedder [51] and TART [1], remain text-only. In contrast, UniIR demonstrates promising cross-dataset generalization in multimodal retrieval, indicating potential advancements for integration into multimodal LLMs.

6. Conclusion

We presented UniIR, a framework to build universal multimodal information retrieval models. This framework enables one unified retriever to follow natural language instruction and accomplish diverse information retrieval tasks across different modalities. We build the M-BEIR benchmark to enable the training and evaluation of UniIR models. We show that our proposed instruction-tuning pipeline can generalize well across different retrieval tasks and domains. However, the existing model performance is still relatively far from perfect indicating ample room for future improvement. We believe that large-scale pre-training algorithms with a stronger vision-language backbone model can build the foundation towards closing this gap and would leave this direction for future exploration.

Author Contributions

CW and YC led the project. The authors had overlapping responsibilities, but the biggest contributions from each author were as follows:

• CW co-designed the experiment and, in particular, was

responsible for collecting and processing the M-BEIR dataset, creating the codebase for the dataset utilities and UniIR models, creating the pipeline for training and evaluating retriever models as well as FAISS retrieval, conducting experiments for the main results and ablation studies, and writing the paper.

- YC co-designed the experiment and, in particular, was responsible for designing the zero-shot model's codebase, conducting experiments for the zero-shot baselines, and writing the paper.
- HC contributed to the code, particularly the BLIP model.
- GZ and JF contributed to the final writeup.
- HH, AR and WC contributed advice on the project, as well as feedback on writing and presentation.

Acknowledgments

Yang Chen and Alan Ritter are supported by the NSF (IIS-2052498) and by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via the HIATUS Program contract #2022-22072200004. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies, either expressed or implied, of NSF, ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for governmental purposes notwithstanding any copyright annotation therein.

References

[1] Akari Asai, Timo Schick, Patrick Lewis, Xilun Chen, Gautier Izacard, Sebastian Riedel, Hannaneh Hajishirzi, and Wen-tau Yih. Task-aware retrieval with instructions. *Findings of ACL*, 2022. 5, 9

[2] Akari Asai, Sewon Min, Zexuan Zhong, and Danqi Chen. Retrieval-based language models and applications. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 6: Tutorial Abstracts)*, pages 41–46, 2023. 1

[3] Andreas Blattmann, Robin Rombach, Kaan Oktay, Jonas Müller, and Björn Ommer. Semi-parametric neural image synthesis. In *Advances in Neural Information Processing Systems*, 2022. 1, 9

[4] Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, et al. Improving language models by retrieving from trillions of tokens. In *International conference on machine learning*, pages 2206–2240. PMLR, 2022. 9

[5] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 18392–18402, 2023. 9

[6] Yingshan Chang, Mridu Narang, Hisami Suzuki, Guihong Cao, Jianfeng Gao, and Yonatan Bisk. Webqa: Multihop and multimodal qa. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16495–16504, 2022. 3, 5, 2, 14

[7] Soravit Changpinyo, Jordi Pont-Tuset, Vittorio Ferrari, and Radu Soricut. Telling the what while pointing to the where: Multimodal queries for image retrieval. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 12136–12146, 2021. 9

[8] Wenhu Chen, Hexiang Hu, Xi Chen, Pat Verga, and William W Cohen. Murag: Multimodal retrieval-augmented generator for open question answering over images and text. *Proceedings of Conference on Empirical Methods in Natural Language Processing*, 2022. 9

[9] Wenhu Chen, Hexiang Hu, Chitwan Saharia, and William W Cohen. Re-imagen: Retrieval-augmented text-to-image generator. *The International Conference on Learning Representations*, 2022. 1, 9

[10] 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? *Proceedings of Conference on Empirical Methods in Natural Language Processing*, 2023. 3, 5, 15

[11] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*, 2022. 9

[12] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned language models. *arXiv preprint arXiv:2210.11416*, 2022. 5, 9

[13] Wenliang Dai, Junnan Li, Dongxu Li, Anthony Meng Huat Tiong, Junqi Zhao, Weisheng Wang, Boyang Albert Li, Pascale Fung, and Steven C. H. Hoi. Instructblip: Towards general-purpose vision-language models with instruction tuning. *Advances in neural information processing systems*, 2023. 9

[14] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *The International Conference on Learning Representations*, 2020. 6, 1

[15] Stephanie Fu*, Netanel Tamir*, Shobhita Sundaram*, Lucy Chai, Richard Zhang, Tali Dekel, and Phillip Isola. Dreamsim: Learning new dimensions of human visual similarity using synthetic data. *Advances in neural information processing systems*, 2023. 3, 4, 2, 15

[16] Rohit Girdhar, Alaaeldin El-Nouby, Zhuang Liu, Mannat Singh, Kalyan Vasudev Alwala, Armand Joulin, and Ishan Misra. Imagebind: One embedding space to bind them all. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15180–15190, 2023. 9

[17] Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. Retrieval augmented language model pre-training. In *International conference on machine learning*, pages 3929–3938. PMLR, 2020. 9

[18] Xintong Han, Zuxuan Wu, Phoenix X. Huang, Xiao Zhang, Menglong Zhu, Yuan Li, Yang Zhao, and Larry S. Davis. Automatic spatially-aware fashion concept discovery. In *Proceedings of the IEEE International Conference on Computer Vision*, 2017. 3, 4, 5, 2, 14

[19] Hexiang Hu, Yi Luan, Yang Chen, Urvashi Khandelwal, Mandar Joshi, Kenton Lee, Kristina Toutanova, and Ming-Wei Chang. Open-domain visual entity recognition: Towards recognizing millions of wikipedia entities. *Proceedings of the IEEE International Conference on Computer Vision*, 2023. 2, 3, 4, 15

[20] Ziniu Hu, Ahmet Iscen, Chen Sun, Zirui Wang, Kai-Wei Chang, Yizhou Sun, Cordelia Schmid, David A Ross, and Alireza Fathi. Reveal: Retrieval-augmented visual-language pre-training with multi-source multimodal knowledge memory. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23369–23379, 2023. 9

[21] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, 2021. 1

[22] Gautier Izacard and Édouard Grave. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 874–880, 2021. 9

[23] Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. Few-shot learning with retrieval augmented language models. *Journal of Machine Learning Research*, 2023. 9

[24] Aashi Jain, Mandy Guo, Krishna Srinivasan, Ting Chen, Sneha Kudugunta, Chao Jia, Yinfei Yang, and Jason Baldridge. Mural: multimodal, multitask retrieval across languages. *Findings of the Association for Computational Linguistics: EMNLP*, 2021. 9

[25] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International Conference on Machine Learning*, pages 4904–4916. PMLR, 2021. 9

[26] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547, 2019. 6

[27] Andrej Karpathy and Li Fei-Fei. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3128–3137, 2015. 9, 2

[28] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, pages 5583–5594. PMLR, 2021. 9

[29] Kenton Lee, Ming-Wei Chang, and Kristina Toutanova. Latent retrieval for weakly supervised open domain question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 6086–6096, 2019. 9

[30] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474, 2020. 9

[31] Dongxu Li, Junnan Li, Hung Le, Guangsen Wang, Silvio Savarese, and Steven C.H. Hoi. LAVIS: A one-stop library for language-vision intelligence. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, 2023. 1

[32] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems*, 34:9694–9705, 2021. 9

[33] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900. PMLR, 2022. 2, 5, 6, 1

[34] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *International Conference on Machine Learning*, 2023. 6, 1

[35] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014. 1, 3, 4, 14

[36] Xi Victoria Lin, Xilun Chen, Mingda Chen, Weijia Shi, Maria Lomeli, Rich James, Pedro Rodriguez, Jacob Kahn, Gergely Szilvasy, Mike Lewis, et al. Ra-dit: Retrieval-augmented dual instruction tuning. *arXiv preprint arXiv:2310.01352*, 2023. 9

[37] Fuxiao Liu, Yinghan Wang, Tianlu Wang, and Vicente Ordonez. Visual news: Benchmark and challenges in news image captioning. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2021. 3, 4, 2, 14

[38] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 2023. 9

[39] Siqi Liu, Weixi Feng, Wenhu Chen, and William Yang Wang. Edis: Entity-driven image search over multimodal web content. *Proceedings of Conference on Empirical Methods in Natural Language Processing*, 2023. 1, 2, 3, 4, 14

[40] Zheyuan Liu, Cristian Rodriguez-Opazo, Damien Teney, and Stephen Gould. 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, 2021. 3, 4, 2, 15

[41] Zhenghao Liu, Chenyan Xiong, Yuanhuiyi Lv, Zhiyuan Liu, and Ge Yu. Universal vision-language dense retrieval: Learning a unified representation space for multi-modal retrieval. In *The Eleventh International Conference on Learning Representations*, 2022. 2

[42] Man Luo, Zhiyuan Fang, Tejas Gokhale, Yezhou Yang, and Chitta Baral. End-to-end knowledge retrieval with multimodal queries. *Annual Meeting of the Association for Computational Linguistics*, 2023. 9

[43] Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. Cross-task generalization via natural language crowdsourcing instructions. In *Annual Meeting of the Association for Computational Linguistics*, 2021. 9

[44] OpenAI. Gpt-4 technical report. *ArXiv*, abs/2303.08774, 2023. 9

[45] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35: 27730–27744, 2022. 9

[46] Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In *Proceedings of the IEEE international conference on computer vision*, pages 2641–2649, 2015. 9

[47] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PMLR, 2021. 2, 5, 6, 1

[48] Ori Ram, Yoav Levine, Itay Dalmedigos, Dor Muhlgay, Amnon Shashua, Kevin Leyton-Brown, and Yoav Shoham. In-context retrieval-augmented language models. *Transactions of the Association for Computational Linguistics*, 2023. 1

[49] Shelly Sheynin, Oron Ashual, Adam Polyak, Uriel Singer, Oran Gafni, Eliya Nachmani, and Yaniv Taigman. kNN-diffusion: Image generation via large-scale retrieval. In *The Eleventh International Conference on Learning Representations*, 2023. 1, 9

[50] Amit Singhal et al. Modern information retrieval: A brief overview. *IEEE Data Eng. Bull.*, 24(4):35–43, 2001. 1

[51] Hongjin Su, Weijia Shi, Jungo Kasai, Yizhong Wang, Yushi Hu, Mari Ostendorf, Wen-tau Yih, Noah A. Smith, Luke Zettlemoyer, and Tao Yu. One embedder, any task: Instruction-finetuned text embeddings. In *Findings of the Association for Computational Linguistics: ACL 2023*, 2023. 9

[52] Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. Beir: A heterogenous benchmark for zero-shot evaluation of information retrieval models. *Advances in neural information processing systems*, 2021. 9

[53] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023. 9

[54] Boxin Wang, Wei Ping, Lawrence McAfee, Peng Xu, Bo Li, Mohammad Shoeybi, and Bryan Catanzaro. Instructretro: Instruction tuning post retrieval-augmented pretraining. *arXiv preprint arXiv:2310.07713*, 2023. 9

[55] Jason Wei, Maarten Bosma, Vincent Y Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. *The International Conference on Learning Representations*, 2021. 5, 9

[56] Hui Wu, Yupeng Gao, Xiaoxiao Guo, Ziad Al-Halah, Steven Rennie, Kristen Grauman, and Rogerio Feris. 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, 2021. 3, 4, 5, 2, 15

[57] Zhiyang Xu, Ying Shen, and Lifu Huang. Multiinstruct: Improving multi-modal zero-shot learning via instruction tuning. *Annual Meeting of the Association for Computational Linguistics*, 2023. 9

[58] Michihiro Yasunaga, Armen Aghajanyan, Weijia Shi, Richard James, Jure Leskovec, Percy Liang, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. Retrieval-augmented multimodal language modeling. 2023. 1, 9

[59] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. *Proceedings of the IEEE International Conference on Computer Vision*, 2023. 6, 1

[60] Kai Zhang, Lingbo Mo, Wenhu Chen, Huan Sun, and Yu Su. Magicbrush: A manually annotated dataset for instruction-guided image editing. *Advances in neural information processing systems*, 2023. 9

UniIR: Training and Benchmarking Universal Multimodal Information Retrievers

Supplementary Material

6.1. Visualization of Retrieval Results

We present example retrieval results for all 10 datasets in the M-BEIR benchmark, shown in Figures 7 through 16. For each dataset, we show one example query and the top 5 candidates retrieved from M-BEIR (5.6 M) by UniIR (CLIP_{SF}), UniIR (BLIP_{FF}), and Multi-task fine-tuned baseline retrievers Multi-task (CLIP_{SF}) and Multi-task (BLIP_{FF}). Our visualization further demonstrates the findings that we have discussed in Section 4.

Instruction-tuning improves retrieval on M-BEIR. For most of the tasks, the multi-task baselines retrieved candidates from undesired modalities. As demonstrated in Figures 7, 8, 9 and 10, the multi-task baselines only retrieved text candidates that had a similar meaning to the query and failed to understand the intent of the task. For instance, in the VisualNews example 7, the multi-task baselines are distracted by text that contains the entity names "Obama" and "Gates". Similarly, in the WebQA example 11, baseline models are distracted by text that contains "Saint Peter". On the other hand, UniIR models were able to accurately fetch candidates in the desired modalities.

Single-modality retrieval task. Even without instructions, the baseline models can perform well in single-modality retrieval tasks, such as similar image retrieval in Figure 12. This is consistent with the results presented in Section 4, as these tasks do not require a model to understand intent.

Vision and language composed image retrieval task. In Table 2, it is evident that UniIR (BLIP_{FF}) outperforms UniIR (CLIP_{SF}) in the composed image retrieval task (task 7). The composed image retrieval task generally involves modifying a reference image according to a textual description. Our hypothesis is that the "cross-attention" layer in (BLIP_{FF}) can effectively merge the image and text embeddings, making it better suited for this task. Our presented examples 13 and 14 indicate that models based on BLIP_{FF} can accurately retrieve the target candidate.

InfoSeek and OVEN tasks. InfoSeek task is more challenging as compared to the OVEN task, which mainly focuses on visual entity recognition. Infoseek queries require details from Wikipedia articles. As illustrated in Example 16, only UniIR (BLIP_{FF}) successfully retrieved the target candidate in the top five results. Although UniIR (CLIP_{SF}) retrieved 5 related candidates that correctly identified the lake in the query image, none of them contained the desired answer.

6.2. Experiment Details

Zero-shot SoTA Retriever. We utilize pre-trained vision-language models such as CLIP (L-14)[47], SigLIP (L)[59], BLIP (L) [33], and BLIP2 [34] as our baseline feature extractors. We use score-level fusion with $w_1 = w_2 = 1$ to fuse multimodal features (i.e., element-wise addition). For task 8 (OVEN and InfoSeek), which contains image-text queries and candidates, we exclusively use the image for the input query and candidate, as we found it achieves better performance in preliminary studies than using both modalities. In addition, we use the Wikipedia page title as the candidate with the prompt "`a picture of [title]`" as we found this approach gives better zero-shot performance compared to using 100 tokens from the Wikipedia page as the candidate. Although BLIP and BLIP2 inherently support multimodal feature extraction, these features were not pre-trained specifically for retrieval tasks. Therefore, we apply the same approach as CLIP to extract features (dim=256). We use the CLIP (`ViT-L-14`), BLIP (`BLIP/models/model_large.pth`) and BLIP2 (`BLIP2/blip2_pretrained.pt`) models from the LAVIS [31] library with the supported `feature_extractor` function. For SigLIP (`timm/ViT-L-16-SigLIP-256`) model, we use the Open-CLIP [21] library.

UniIR Models and Multi-task fine-tuned Models. We use the largest checkpoint, ViT-L14 [14], for all CLIP and BLIP variants. The default image resolution is 224×224, and for UniIR (BLIP_{FF}), we also report the finetuned results with a resolution of 384×384, which is a commonly used setting for BLIP finetuning. Our batch size is 105 for CLIP variants and 115 for BLIP variants on the 224×224 resolution. We train the model for 20 epochs using other hyperparameters as reported in the original implementations. For score fusion methods, we set $w_1 = w_2 = w_3 = w_4 = 1$ by default. CLIP_{FF} uses a 2-layer transformer architecture similar to the T5 Transformer [47], but with only 2 layers and 12 attention heads, with each head having 64 dimensions. BLIP_{FF} follows the original implementation of BLIP. The output from the image encoder is fed into transformer layers of the image-grounded text encoder through cross-attention layers, and then the output is treated as the fused feature. We train all our models using in-batch query-candidate contrastive loss [47] to maximize the positive pair similarity while minimizing negative pair similarity. All of our experiments are conducted on a single node with 8 H100 GPUs.

6.3. Data Collection

M-BEIR Format. Each query instance in the M-BEIR dataset includes a query $\mathbf{q}$, a list of relevant(positive) candidate data $\mathbf{c}^+$, and a list of potentially available irrelevant(negative) candidate data $\mathbf{c}^-$. In addition, a human-authored instruction $q_{\texttt{inst}}$ is provided with $\mathbf{q}$ to specify the intent of the retrieval task. It's important to note that every M-BEIR query instance has at least one positive candidate data and possibly no negative candidate data.

VisualNews. We follow the preprocessing pipeline outlined in the Visual News dataset [37] and randomly sampled 200K, 40K, and 40K image-caption pairs for our training, validation, and test sets, respectively. Then, converted them into M-BEIR format. As a result, we have 100K instances for task 1 $q_t \rightarrow c_i$ and 100K instances for task 4 $q_i \rightarrow c_t$ in the training set. We used all the images and captions in the Visual News dataset as the initial candidate pool, which amounted to a total of 2.5M candidates. We then trimmed down the candidate pool to 1M, ultimately arriving at a final M-BEIR$_{\text{local}}$ pool of 500K text and 500K images.

Fashion200K. The Fashion200K dataset [18] consists of image and product description pairs. We use all the available images and descriptions as our M-BEIR$_{\text{local}}$ candidate pool, totalling 260K entries, with 200K images and 60K text descriptions. We randomly selected 30K image-description pairs from the original training set to form our training data. The original test data is evenly divided into a validation and test set. We converted the dataset into M-BEIR format. In total, we have 15K task 1 ($q_t \rightarrow c_i$) instances and 15K task 4 ($q_i \rightarrow c_t$) instances in training set.

MSCOCO. We used the Karpathy split [27] for MSCOCO, which contains 113K/5K/5K images for train/validation/test. We directly converted MSCOCO into M-BEIR format. This results in 113K task 4 ($q_i \rightarrow c_t$) instance and 566K task 1 ($q_t \rightarrow c_i$) instances for the training set. We then trimmed the task 1 instance down to 100k. When evaluating models on the MSCOCO test split, we use all the images and captions from the original test set to construct the test split M-BEIR$_{\text{local}}$ candidate pool, which has 25K text and 5K images.

WebQA. We followed the full-scale retrieval setting in WebQA [6] and constructed the 940K M-BEIR$_{\text{local}}$ candidate pool. This pool contains 400K pairs of (image, text) and 540K text-only candidates. We convert the WebQA dataset into the M-BEIR format by using questions as queries. We randomly sample 5000 instances from the original training set as our test set. As a result, the M-BEIR WebQA training set consists of 15K task 2 instances of $q_i \rightarrow c_t$ and 15K task 3 instances of $q_i \rightarrow (c_i, c_t)$.

EDIS. We followed the full candidate retrieval setting in EDIS [39] and use all the 1M image-headline candidates in EDIS as the M-BEIR$_{\text{local}}$ candidate pool. We directly convert the original train/val/test split of 26K/3.2K/3.2K instances into M-BEIR format, which involves using the caption as the query and the image-headline pair as the positive candidate. This process results in 26K task 3 $q_i \rightarrow (c_i, c_t)$ instances in the training set.

NIGHTS. To convert the 20K triplets in the 2AFC NIGHTS dataset [15] into M-BEIR format, we used the reference image as the query. For the positive candidate, we selected the image target aligned with human judgment, and for the negative candidate, we chose the other image target that disagreed with human judgment. This results in 16K/2K/2K task 5 $q_i \rightarrow c_i$ instances for train/val/test split. We use all the 40K images in the NIGHTS dataset as the M-BEIR$_{\text{local}}$ task-specific candidate pool.

FashionIQ. The FashionIQ dataset [56] consists of pairs of reference and target images alongside two sets of captions. However, since the captions are not detailed enough, we follow the approach of [40] and concatenated them into a single caption. To convert the FashionIQ dataset into M-BEIR format, we use the reference image and concatenated caption as a query and the target image as a positive candidate. All images in FashionIQ serve as the candidate pool for M-BEIR$_{\text{local}}$. We randomly sampled 1.7K instances from the converted training set as our validation set and used the original validation set as the test set.

CIRR. The CIRR dataset [40], comprises pairs of reference and target images, along with a modification sentence that describes the changes made. We consider the reference image and modification sentence as the query and the target image as a positive candidate for a task 7 $(q_i, q_t) \rightarrow c_i$ instance. We use all the images in the CIRR dataset as the M-BEIR$_{\text{local}}$ candidate pool. For validation purposes, we randomly select 2K instances from the training set, and the original validation set is used as the test set.

OVEN. The OVEN dataset [19] has instances that include an image and a visual recognition text question. Additionally, it has a related image(potentially empty) and its corresponding text description as the target candidate. In order to convert this dataset into M-BEIR format, we treat the image and text question pair as a query, and the image and text description pair as a positive candidate. The text description in OVEN [19] is simply the title of the Wikipedia subject. However, in order to create the candidate text, we concatenate the Wikipedia title with the first 100 tokens of its summary. This allows for a more comprehensive understanding of the text description. This results in 4M training instances for task 6 $(q_i, q_t) \rightarrow c_t$ and 700K instances for task 8 $(q_i, q_t) \rightarrow (c_i, c_t)$. We trimmed the instances down to 120K each. We also trimmed down the original 6M candidates pool to 1M and adopted it as M-BEIR$_{\text{local}}$.

InfoSeek. The Infoseek dataset [10] comprises image and text question pairs, along with an associated image (which may be empty) and a corresponding text description that contains the answer. To convert this dataset to M-BEIR format, we treat the image and text question pair as a query, and the image and text description pair as a positive candidate. However, the text description in Infoseek [10] is an entire Wikipedia article with thousands of tokens. To create a proper retrieval task, we split the Wikipedia article into snippets of 100 tokens and use the snippet which contains the exact answer as the positive candidate, and the rest as negative candidates. This results in 400K training instances for Task 6 $(q_i, q_t) \rightarrow c_t$, and 300K instances for Task 8 $(q_i, q_t) \rightarrow (c_i, c_t)$. We trimmed the instances down to 140K each. We also trimmed down the original 6M candidates pool to 1M and adopted it as M-BEIR$_{\text{local}}$.

We removed distorted text or images and resized the image to 256 pixels for the shorter dimension in all the datasets.

M-BEIR Candidates Pool. The M-BEIR (5.6M) candidate pool is created by combining M-BEIR$_{\text{local}}$ pools from all 10 datasets.

6.4. Additional Experiment Analysis

Zero-shot Retriever Results. In Table 6, we benchmark zero-shot retrieval models on each task and found that SigLIP performed the best on average. Interestingly, while the BLIP model performed on par with SigLIP or even better than CLIP on the MSCOCO dataset, which focuses on common objects, it underperformed significantly on other datasets in domains such as news, fashion, and Wikipedia. However, we observed that SigLIP significantly outperformed all other models on Fashion200K (36.5) and FashionIQ (14.4), while CLIP had a clear advantage on news datasets such as VisualNews (43.3). This suggests that covering a wide range of visual domains during pre-training, rather than focusing on a single domain, is crucial for achieving robust retrieval. M-BEIR presents a comprehensive evaluation benchmark to serve this purpose.

Held-out Dataset Generalization. In Table 7 and 8, we report full results of each model used in Figure 5 and 6 on M-BEIR and M-BEIR$_{\text{local}}$, respectively.

		ZS	ZS	ZS	Multi.	UniIR	ZS	Multi.	UniIR
Task	**Dataset**	SigLIP	BLIP2	CLIP	CLIP$_{\text{SF}}$	CLIP$_{\text{SF}}$	BLIP	BLIP$_{\text{FF}}$	BLIP$_{\text{FF}}$
2.	WebQA	34.1	35.2	32.1	35.0	51.8	38.1	36.7	37.1
3.	WebQA	2.2	0.0	5.5	10.2	20.8	0.0	30.4	35.3
6.	OVEN	0.0	0.0	0.0	8.9	6.2	0.0	10.9	12.0
7.	CIRR	7.1	6.2	5.4	14.1	16.9	7.4	15.7	34.1
8.	OVEN	27.2	13.8	24.5	34.4	43.7	10.1	20.5	24.9
-	Average	14.1	11.0	13.5	20.5	27.9	11.1	22.8	28.7

Table 7. **Held-out dataset generalization** experiments (Recall@5) on M-BEIR: we train a Multi-task (Multi.) and a UniIR model on 7 held-in datasets and test on 3 held-out datasets (WebQA, OVEN, CIRR).

		ZS	ZS	ZS	Multi.	UniIR	ZS	Multi.	UniIR
Task	**Dataset**	SigLIP	BLIP2	CLIP	CLIP$_{\text{SF}}$	CLIP$_{\text{SF}}$	BLIP	BLIP$_{\text{FF}}$	BLIP$_{\text{FF}}$
2.	WebQA	39.8	38.6	36.3	37.0	54.5	44.9	41.3	41.3
3.	WebQA	43.5	24.5	45.1	46.8	47.7	20.3	51.7	50.3
6.	OVEN	29.7	12.2	24.2	21.6	25.0	16.1	14.2	16.5
7.	CIRR	22.7	11.8	13.2	19.7	21.4	10.6	29.4	37.1
8.	OVEN	41.7	27.3	38.8	50.3	52.8	27.4	25.6	30.7
-	Average	35.5	22.9	31.5	35.1	40.3	23.8	32.4	35.2

Table 8. **Held-out dataset generalization** experiments (Recall@5) on M-BEIR$_{\text{local}}$: we train a Multi-task (Multi.) and a UniIR model on 7 held-in datasets and test on 3 held-out datasets (WebQA, OVEN, CIRR).

Dataset	Domain	Task	Query Instruction $q_{\texttt{inst}}$	Query Image q_i	Query Text q_t
VisualNews	News	1. $q_t \rightarrow c_i$	Based on the caption, provide the most fitting image for the news story.	-	Robert Gates and President Obama salute during a farewell ceremony for Gates as outgoing defense secretary Gates served in that position under Obama and President George W Bush He was known as a savvy manager of the bureaucracy and a ruthlessly efficient defense secretary In his new memoir Gates asserts that Obama had more than doubts about his course in Afghanistan.

Model	Rank 1 c_i	Rank 2 c_i	Rank 3 c_i	Rank 4 c_i	Rank 5 c_i
UniIR (CLIP_{SF}) ✓inst	✗ 	✗ 	✗ 	✗ 	✓ 
multi-task (CLIP_{SF}) ✗inst	✗ Robert Gates and President Obama salute during a farewell ceremony for Gates as outgoing defense secretary Gates served in that position under Obama and President George W Bush He was known as a savvy manager of the bureaucracy and a ruthlessly efficient defense secretary In his new memoir Gates asserts that Obama had more than doubts about his course in Afghanistan.	✗ Former Defense Secretary Robert Gates acknowledges applause as President Obama speaks at Gates s farewell ceremony at the Pentagon in June 2011.	✗ Former Defense secretary Robert Gates is seen with President Obama.	✗ President Obama and Robert Gates.	✗ Obama presents Gates with the Presidential Medal of Freedom The citation said Gates had selflessly dedicated his life to ensuring the security of the American people Gates cultivated an eventempered image in his years in government.
UniIR (BLIP_{FF}) ✓inst	✗ 	✗ 	✓ 	✗ 	✗ 
multi-task (BLIP_{FF}) ✗inst	✗ Robert Gates and President Obama salute during a farewell ceremony for Gates as outgoing defense secretary Gates served in that position under Obama and President George W Bush He was known as a savvy manager of the bureaucracy and a ruthlessly efficient defense secretary In his new memoir Gates asserts that Obama had more than doubts about his course in Afghanistan.	✗ Former Defense Secretary Robert Gates acknowledges applause as President Obama speaks at Gates s farewell ceremony at the Pentagon in June 2011.	✗ 	✗ Former Defense secretary Robert Gates is seen with President Obama.	✗ 
Zero-shot (BLIP2) ✗inst	✗ Robert Gates and President Obama salute during a farewell ceremony for Gates as outgoing defense secretary Gates served in that position under Obama and President George W Bush He was known as a savvy manager of the bureaucracy and a ruthlessly efficient defense secretary In his new memoir Gates asserts that Obama had more than doubts about his course in Afghanistan.	✗ Former Defense Secretary Robert Gates acknowledges applause as President Obama speaks at Gates s farewell ceremony at the Pentagon in June 2011.	✗ Former US president George W Bush center shakes hands with his father former President George HW Bush as they attend the opening ceremony of the George W Bush Presidential Center on April 25 2013 in Dallas Texas.	✗ President Obama right and Senior Chief Special Warfare Operator Edward Byers Jr arrive for a ceremony at the White House where Byers received the Medal of Honor for his courageous actions in Afghanistan in 2012.	✗ President Obama right and former president George W Bush arrive for a wreathlaying ceremony for victims of the 1998 bombing at the US Embassy in Dar es Salaam Tanzania.

Figure 7. Visualization of the Top 5 Candidates Retrieved from M-BEIR (5.6M candidates) by Different Models for a VisualNews Query

Dataset	Domain	Task	Query Instruction $q_{\texttt{inst}}$	Query Image q_i	Query Text q_t
MSCOCO	Common	1. $q_t \rightarrow c_i$	Show me an image that best captures the following common scene description.	-	An antilope is eating grass in between two zebra.

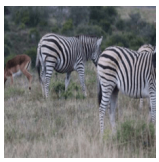
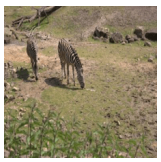
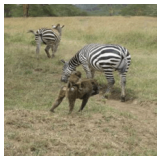
Model	Rank 1 c_i	Rank 2 c_i	Rank 3 c_i	Rank 4 c_i	Rank 5 c_i
UniIR (CLIP_{SF}) ✓inst	✗ 	✓ 	✗ 	✗ 	✗ 
multi-task (CLIP_{SF}) ✗inst	✗ An antilope is eating grass in between two zebra.	✗ Two zebras and another animal grazing in the grass.	✗ A couple of zebras stand next to a antelope.	✗ Monkey standing behind two zebras as they graze.	✗ Three zebras and two other animals grazing.
UniIR (BLIP_{FF}) ✓inst	✗ 	✗ 	✗ 	✗ 	✗ 
multi-task (BLIP_{FF}) ✗inst	✗ An antilope is eating grass in between two zebra.	✗ Two zebras and another animal grazing in the grass.	✗ A couple of zebras graze on some grass.	✗ Two zebras are feeding on the grass by themselves.	✗ 
Zero-shot (BLIP2) ✗inst	✗ An antilope is eating grass in between two zebra.	✗ A couple of zebra's leaned over eating grass in a field.	✗ A zebra eating green grassy items for his lunch.	✗ A zebra eating long clipped grass on the ground.	✗ A zebra standing in a field grazing as another looks alert.

Figure 8. Visualization of the Top 5 Candidates Retrieved from M-BEIR (5.6M candidates) by Different Models for an MSCOCO Query

Dataset	Domain	Task	Query Instruction $q_{\texttt{inst}}$	Query Image q_i	Query Text q_t
Fashion200K	Fashion	1. $q_t \rightarrow c_i$	Based on the following fashion description, retrieve the best matching image.	-	Natural pebbled chevron alpaca drape front jacket.

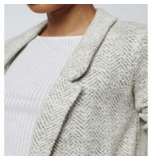
Model	Rank 1 c_i	Rank 2 c_i	Rank 3 c_i	Rank 4 c_i	Rank 5 c_i
UniIR (CLIP_{SF}) ✓inst	 ✗	 ✗	 ✗	 ✗	 ✗
multi-task (CLIP_{SF}) ✗inst	✗ Natural pebbled chevron alpaca drape front jacket.	✗ Natural drape front open jacket.	✗ Multicolor drape front jacket.	✗ Multicolor crepe drape front jacket.	✗ Multicolor drape front wool jacket.
UniIR (BLIP_{FF}) ✓inst	 ✗	 ✗	 ✗	 ✗	 ✓
multi-task (BLIP_{FF}) ✗inst	✗ Natural pebbled chevron alpaca drape front jacket.	✗ Beige textured terry draped jacket.	 ✗	 ✗	✗ Multicolor cashmere shawl collar jacket.
Zero-shot (BLIP2) ✗inst	✗ Natural pebbled chevron alpaca drape front jacket.	✗ Beige corded zigzag knit jacket.	✗ Natural drape front open jacket.	✗ Beige textured terry draped jacket.	✗ Natural vegan sherpa drape jacket.

Figure 9. Visualization of the Top 5 Candidates Retrieved from M-BEIR (5.6M candidates) by Different Models for a Fashion200K Query

Dataset	Domain	Task	Query Instruction $q_{\texttt{inst}}$	Query Image q_i	Query Text q_t
EDIS	News	3. $q_t \rightarrow (c_i, c_t)$	Find a news image that matches the provided caption.	-	Barack Obama with Germany's chancellor Angela Merkel at the Brandenburg Gate Berlin on 19 June.

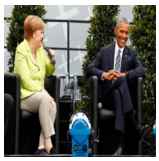
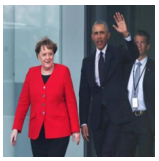
Model	Rank 1 (c_i, c_t)	Rank 2 (c_i, c_t)	Rank 3 (c_i, c_t)	Rank 4 (c_i, c_t)	Rank 5 (c_i, c_t)
UniIR (CLIP_{SF}) ✓inst	 ✗ Obama Speaks at a Berlin Event With Angela Merkel.	 ✓ Obama's Berlin visit to coincide with Trump in Brussels - Barack Obama.	 ✗ Obama meets Germany's Merkel at chancellery in Berlin.	 ✓ President Obama Speaks to the People of Berlin from the Brandenburg Gate.	 ✓ When Barack Obama visited Berlin two years ago, he charmed a city.
multi-task (CLIP_{SF}) ✗inst	✗ President Obama stands next to German Chancellor Angela Merkel in front of Brandenburg Gate in Berlin on June 19.	✗ President Obama and German Chancellor Angela Merkel in 2011.	 ✓ When Barack Obama visited Berlin two years ago, he charmed a city.	✗ Barack Obama with German Chancellor Angela Merkel at the G20 summit in November.	✗ US president Barack Obama at the Brandenburg Gate.
UniIR (BLIP_{FF}) ✓inst	 ✓ When Barack Obama visited Berlin two years ago, he charmed a city.	 ✗ Obama Gets Warm Welcome in Berlin, Praises Merkel.	 ✗ Obama to give speech at Berlin's Brandenburg Gate.	 ✗ As President Obama's team grapples with complaints from other countries about U.	 ✓ President Obama Speaks to the People of Berlin from the Brandenburg Gate.
multi-task (BLIP_{FF}) ✗inst	✗ President Obama stands next to German Chancellor Angela Merkel in front of Brandenburg Gate in Berlin on June 19.	✗ President Obama and German Chancellor Angela Merkel in 2011.	 ✓ When Barack Obama visited Berlin two years ago, he charmed a city.	✗ Angela Merkel greets Barack Obama May 2012 Merkel Finally the Germans get to make love not war.	✗ US President Barack Obama waves next to German Chancellor Angela Merkel before they deliver speeches to invited guests in front of the Brandenburg Gate at Pariser Platz in Berlin on June 19 2013 during the official visit of the US President Barack Obama walks in John F Kennedy s footsteps on his first visit to Berlin as US president but encounter a more powerful...
Zero-shot (BLIP2) ✗inst	✗ President Obama stands next to German Chancellor Angela Merkel in front of Brandenburg Gate in Berlin on June 19.	✗ US President Barack Obama waves next to German Chancellor Angela Merkel before they deliver speeches to invited guests in front of the Brandenburg Gate at Pariser Platz in Berlin on June 19 2013 during the official visit of the US President Barack Obama walks in ...	✗ Barack Obama with German chancellor Angela Merkel at the G20 summit in November.	✗ BERLIN GERMANY JUNE 19 US President Barack Obama meets German Chancellor Angela Merkel for bilateral talks at the Chancellery on June 19 2013 in Berlin Germany Obama is visiting Berlin for the first time during his presidency and his speech at the Brandenburg Gate...	✗ President Obama walks with Germany s Chancellor Angela Merkel in St Petersburg Russia on Sept 6.

Figure 10. Visualization of the Top 5 Candidates Retrieved from M-BEIR (5.6M candidates) by Different Models for an EDIS Query

Dataset	Domain	Task	Query Instruction $q_{\texttt{inst}}$	Query Image q_i	Query Text q_t
WebQA	Wikipedia	3. $q_t \rightarrow (c_i, c_t)$	Find a Wikipedia image that answers this question.	-	What supports the statues standing on top of a structure at the burial site of Saint Peter?

Model	Rank 1 (c_i, c_t)	Rank 2 (c_i, c_t)	Rank 3 (c_i, c_t)	Rank 4 (c_i, c_t)	Rank 5 (c_i, c_t)
UniIR ($CLIP_{SF}$) ✓inst	 ✗ Statue of St. Peter in St. Peter's Square at the Vatican.	 ✗ Symbol and statues in St. Peter symbol and statues in St. Peter.	 ✗ St-peter St. Peter.	 ✗ Statue in St. Peter's Basilica.	 ✗ The floor above Saint Peter's tomb.
multi-task ($CLIP_{SF}$) ✗inst	✗ Saint Peter's tomb is supposedly directly below the high altar of the basilica. For this reason, many popes have been interred at St. Peter's since the Early Christian period.	✗ Since the crucifixion and burial of Saint Peter in 64 AD, the spot was thought to be the location of the tomb of Saint Peter, where there stood a small shrine. With its increasing prestige the church became richly decorated with statues, furnishings and elaborate chandeliers, and side tombs and altars were continuously added.	✗ Saint Peter's tomb is a site under St. Peter's Basilica that includes several graves and a structure said by Vatican authorities to have been built to memorialize the location of Saint Peter's grave.	✗ Catholic tradition holds that the basilica is the burial site of Saint Peter, chief among Jesus 's apostles and also the first Bishop of Rome (Pope). Saint Peter's tomb is supposedly directly below the high altar of the basilica. For this reason, many popes have been interred at St. Peter's since the Early Christian period.	 ✗ Symbol and statues in St. Peter symbol and statues in St. Peter.
UniIR ($BLIP_{FF}$) ✓inst	 ✗ Statue of St. Peter in St. Peter's Square at the Vatican.	 ✗ St-peter St. Peter.	 ✓ Saint Peter's Basilica, Vatican City, Rome Burial site of Saint Peter, the "First Pope"!	 ✗ Symbol and statues in St. Peter symbol and statues in St. Peter.	✗ St. Peter's Basilica, believed to be the burial site of St. Peter, seen from the River Tiber.
multi-task ($BLIP_{FF}$) ✗inst	 ✗ Statue of St. Peter in St. Peter's Square at the Vatican.	 ✗ Symbol and statues in St. Peter symbol and statues in St. Peter.	 ✗ A Sculpture Featuring Jesus, Saint Paul, Saint Peter and a Pope, St. Peter Basilica, Vatican, Italy Editorial Photo.	 ✗ Statue in St. Peter's Basilica.	✗ Saint Peter's tomb is a site under St. Peter's Basilica that includes several graves and a structure said by Vatican authorities to have been built to memorialize the location of Saint Peter's grave.
Zero-shot (BLIP2) ✗inst	✗ Also a bronze statue of the apostle St. Paul was placed on the top platform, to go with that of St. Peter on Trajan's Column (27 October 1588). (Originally the top platform probably had a statue of Marcus Aurelius, but it had been already lost by the 16th century.).	✗ Pope Sixtus V was determined to erect the obelisk in front of St Peter's, of which the nave was yet to be built.	✗ Along the roofline, by contrast, there is a line of statues, many in their own small pavilions, culminating in Saint Mark flanked by six angels in the centre, above a large gilded winged lion (his symbol, and that of Venice).	✗ Saint Peter's tomb is a site under St. Peter's Basilica that includes several graves and a structure said by Vatican authorities to have been built to memorialize the location of Saint Peter's grave. St. Peter's tomb is near the west end of a...	✗ Saint Peter's tomb is a site under St. Peter's Basilica that includes several graves and a structure said by Vatican authorities to have been built to memorialize the location of Saint Peter's grave.

Figure 11. Visualization of the Top 5 Candidates Retrieved from M-BEIR (5.6M candidates) by Different Models for a WebQA Query

Dataset	Domain	Task	Query Instruction $q_{\texttt{inst}}$	Query Image q_i	Query Text q_t
NIGHTS	Common	5. $q_t \rightarrow c_t$	Which everyday image is the most similar to the reference image?		-

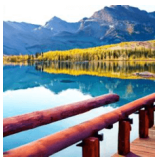
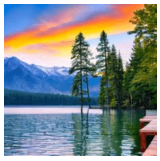
Model	Rank 1 (c_t)	Rank 2 (c_t)	Rank 3 (c_t)	Rank 4 (c_t)	Rank 5 (c_t)
UniIR (CLIP_{SF}) ✓inst	✗ 	✗ 	✗ 	✗ 	✓ 
multi-task (CLIP_{SF}) ✗inst	✓ 	✗ 	✗ 	✗ 	✗ 
UniIR (BLIP_{FF}) ✓inst	✗ 	✗ 	✓ 	✗ 	✗ 
multi-task (BLIP_{FF}) ✗inst	✗ 	✗ 	✓ 	✗ 	✗ 
Zero-shot (BLIP2) ✗inst	✗ 	✗ 	✓ 	✗ 	✗ 

Figure 12. Visualization of the Top 5 Candidates Retrieved from M-BEIR (5.6M candidates) by Different Models for a NIGHTS Query

Dataset	Domain	Task	Query Instruction $q_{\texttt{inst}}$	Query Image q_i	Query Text q_t
FashionIQ	Fashion	7. $(q_i, q_t) \rightarrow c_i$	Find a fashion image that aligns with the reference image and style note.		Is white with dots and a black belt.

Model	Rank 1 c_t	Rank 2 c_t	Rank 3 c_t	Rank 4 c_t	Rank 5 c_t
UniIR (CLIP_{SF}) ✓inst	✗ 	✗ 	✗ 	✗ 	✗ 
multi-task (CLIP_{SF}) ✗inst	✗ 	✗ 	✗ 	✗ 	✗ 
UniIR (BLIP_{FF}) ✓inst	✗ 	✗ 	✓ 	✗ 	✗ 
multi-task (BLIP_{FF}) ✗inst	✗ 	✗ 	✓ 	✗ 	✗ 
Zero-shot (BLIP2) ✗inst	✗ 	✗ 	✗ 	✗ 	✗ 

Figure 13. Visualization of the Top 5 Candidates Retrieved from M-BEIR (5.6M candidates) by Different Models for a FashionIQ Query

Dataset	Domain	Task	Query Instruction $q_{\texttt{inst}}$	Query Image q_i	Query Text q_t
CIRR	Common	7. $(q_i, q_t) \rightarrow c_i$	I'm looking for a similar everyday image with the described changes.		The target photo is of a black and brown dog laying in the grass with a small red ball at his feet.

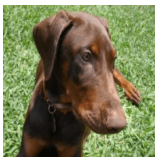
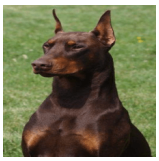
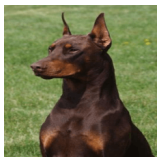
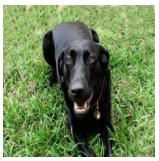
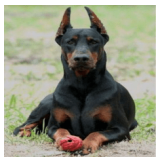
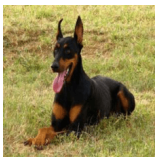
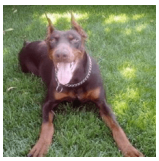
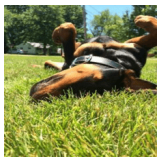
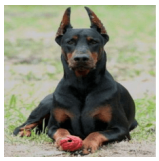
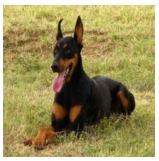
Model	Rank 1 c_t	Rank 2 c_t	Rank 3 c_t	Rank 4 c_t	Rank 5 c_t
UniIR (CLIP_{SF}) ✓inst	✗ 	✗ 	✗ 	✗ 	✗ 
multi-task (CLIP_{SF}) ✗inst	✗ 	✗ 	✗ 	✗ 	✗ 
UniIR (BLIP_{FF}) ✓inst	✓ 	✗ 	✗ 	✗ 	✗ 
multi-task (BLIP_{FF}) ✗inst	✓ 	✗ 	✗ 	✗ 	✗ 
Zero-shot (BLIP2) ✗inst	✗ 	✗ 	✗ 	✗ 	✗ 

Figure 14. Visualization of the Top 5 Candidates Retrieved from M-BEIR (5.6M candidates) by Different Models for a CIRR Query

Dataset	Domain	Task	Query Instruction $q_{\texttt{inst}}$	Query Image q_i	Query Text q_t
OVEN	Wikipedia	8. $(q_i, q_t) \rightarrow (c_i, q_t)$	Determine the Wikipedia image-snippet pair that clarifies the entity in this picture.		What is the name of this place?

Model	Rank 1 (c_i, c_t)	Rank 2 (c_i, c_t)	Rank 3 (c_i, c_t)	Rank 4 (c_i, c_t)	Rank 5 (c_i, c_t)
UniIR ($CLIP_{SF}$) ✓inst	✗  Pará mangroves. other than the city of Belém and its surroundings. The mangroves are used by artisanal fishermen as a source of crabs and wood.Threats include subsistence agriculture and livestock raising, tourism, logging, mining and urban development.	✗  Pará mangroves. The Pará mangroves (NT1427) is an ecoregion along the Atlantic coast of the state of Pará in Brazil.They constitute the western extension of the Maranhão mangroves ecoregion.The mangroves are relatively intact, although they are under...	✓  Itaúnas State Park. by squatters. Activities. As of 2015 the opening hours were from 8:30 to 17:30.The park is suitable for ecotourism, hiking and swimming in the river and the ocean.Visitors may take guided horseback tours, or may explore by kayak, canoe, jeep...	✓  Itaúnas State Park. archaeological sites with traces of human settlements such as chipped stones, indigenous pottery and artifacts from the colonial era.Vegetation includes coastal forest, with fragments of forest endangered in Espírito Santo...	✗  Tree of Life (Bahrain). the list.In October 2010, archaeologists unearthed 500-year-old pottery and other artefacts in the vicinity of the tree. A soil and dendrochronology analysis conducted in the 1990s concluded that the tree was an "Acacia"...
multi-task ($CLIP_{SF}$) ✗ inst	✗  Natal Dunes State Park. The Natal Dunes State Park "Journalist Luiz Maria Alves" (), or simply the Dunes Park () is a state park in the state of Rio Grande do Norte in the Northeast Region of Brazil. It protects an area of dunes and native ...	✗  Cresmina Dune. Cresmina Dune from the north.	✗  Kijal. Kijal (est. pop. (2000 census): 4,375) is a mukim in Kemaman District, Terengganu, Malaysia. The town is well known for the Awana Kijal Golf and Beach Resort, a luxury resort which is owned by the...	✗ Maspalomas Dunes. The Maspalomas Dunes () are sand dunes located on the south coast of the island of Gran Canaria, Province of Las Palmas, in the Canary Islands. A 404 (ha) area of the municipality of San Bartolomé de Tirajana, they have been protected...	✗ Maspalomas Dunes. They were one of the contenders in the 12 Treasures of Spain competition.They were formed by sand from the now subdued marine shelf, when it was laid dry during the last ice age and the wind blew the sand towards the coast of the island...
UniIR ($BLIP_{FF}$) ✓inst	✗  Cronulla sand dunes. context of its setting, intactness, aesthetic qualities and social significance is held in high community esteem. The water bodies formed by sand removal surrounding the Cronulla Sand Dune have been identified ...	✗  Cronulla sand dunes. referred to as the Kurnell sand dune is estimated to be about 15,000 years old. It was formed when the sea reached its present level and began to stabilise, between 9000 and 6000 ...	✗  Tel Tanninim. Tel Tanninim (), in Arabic Tell al-MilÄt (lit. 'Mortar Mound') is an ancient tell (archaeological mound) on the shore of the Mediterranean, near the mouth of Nahal Tanninim ('Crocodiles Stream') ...	✗  Thuli Parks and Wildlife Land. black rhinoceros, lion, Cape wild dog, Namibian cheetah and African leopard, many of which move freely between the protected area and neighbouring Botswana and South Africa ...	✓  Itaúnas State Park. archaeological sites with traces of human settlements such as chipped stones, indigenous pottery and artifacts from the colonial era. Vegetation includes coastal forest, with fragments of forest ...
multi-task ($BLIP_{FF}$) ✗ inst	✗  Cronulla sand dunes. in New South Wales for social, cultural or spiritual reasons.The site has historic and cultural significance for the Aboriginal community. The Cronulla Sand Dune is of cultural heritage and spiritual significance to the La Perouse ...	✗  Cronulla sand dunes. men in the colony. During the 1860s Holt consolidated his landholdings on the Peninsular (which included Captain Cook's landing place). His accumulated holdings represented 5261 (ha) of what is the present day...	✗  Samalayuca Dune Fields. areas of El Paso del Norte in the north and the village of Carrizal in the south. Experiences of travelers in the 1800s.In order to avoid the delay of traveling around the dune fields on the detour, many travelers on the trail...	✗ Maspalomas Dunes. The Maspalomas Dunes () are sand dunes located on the south coast of the island of Gran Canaria, Province of Las Palmas, in the Canary Islands. A 404 (ha) area of the municipality of San BartolomÃ© de Tirajana, they have been protected....	✗ Maspalomas Dunes. The Maspalomas Dunes () They were one of the contenders in the 12 Treasures of Spain competition. They were formed by sand from the now subdued marine shelf, when it was laid dry during the last ice age and the wind blew the sand towards the...
Zero-shot (BLIP2) ✗ inst	✗ 	✗ 	✗  Ad. Imam of Grand Mosque of Mecca, along with the officials of Ministry of Islamic Affairs, Dawah and Guidance also agreed the verse were speaking about ...	✗ 	✗ 

Figure 15. Visualization of the Top 5 Candidates Retrieved from M-BEIR (5.6M candidates) by Different Models for a OVEN Query

Dataset	Domain	Task	Query Instruction $q_{\texttt{inst}}$	Query Image q_i	Query Text q_t
InfoSeek	Wikipedia	8. $(q_i, q_t) \rightarrow (c_i, c_t)$	Retrieve a Wikipedia image-description pair that provides evidence for the question of this image.		Where is the lake outflow to?

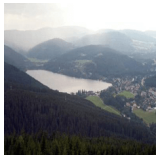
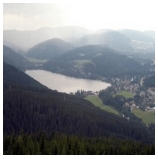
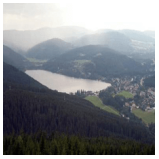
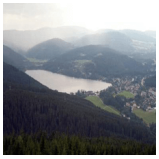
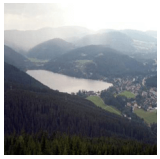
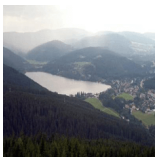
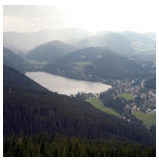
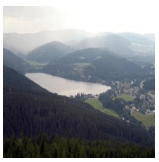
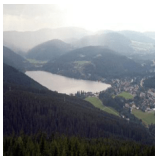
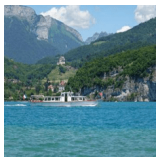
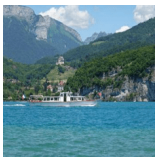
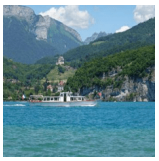
Model	Rank 1 (c_i, c_t)	Rank 2 (c_i, c_t)	Rank 3 (c_i, c_t)	Rank 4 (c_i, c_t)	Rank 5 (c_i, c_t)
UniIR (CLIP_{SF}) ✓inst	 ✗ Titisee. the promenade. Furthermore, plenty of open-air events are organised around the lake each summer.In winter, the 1.2 km long Saig-Titisee toboggan is open. The largest natural ski jump in Germany, the Hochfirstschanze, is also located at ...	 ✗ Titisee. and burbot on the lake bed. In addition there are small fish varieties such as sunbleak, minnow and brook lamprey. Around the lake, grey heron may be seen.The shores of the Titisee are home to two rare types of quillworts, the spring quillwort ...	 ✗ Titisee. could be the origin of the name, even though it no longer occurs by the Titisee. Tourism.At the north shore of the Titisee lies a popular spa town of the same name. Lots of wellness and health hotels are settled there. Spa therapy offers ...	 ✗ Titisee. a tractor with a snowplough was clearing the landing strip of snow when it broke through the ice and sank to the bottom of the lake, taking the driver, Walter Wilde (29), with it. His body was only recovered 2 weeks later. Fauna and flora ...	 ✗ Titisee. great fire. In the 18th century, the clockmaking trade developed in the town to become a major part of the economy.During World War I and shortly thereafter, a dearth of staple foods prevailed. In May 1919 the first municipa ...
multi-task (CLIP_{SF}) ✗inst	 ✗ Wörthersee. ("Coregonus lavaretus"). Etymology.First mentioned as "Werdse" in an 1143 deed, the lake's name originates from the islands in the lake, most notably Maria Wörth, a peninsula since the building of the Glanfurt outlet in 1770...	 ✗ Wörthersee. The lake's water is of a distinctive blue-green colour and transparent. Lake Wörth and its basin in the central Carinthian foothills were largely formed by glaciers during the last ice age. The lake is divided into three basins by several...	 ✗ Titisee. the High Black Forest had been unsettled during the first millennium. Origin of the name.There are various theories about the origin of the unusual name "Titisee": In the Alemannic dialect "Teti" means "little child" or "baby"...	 ✗ Titisee. and burbot on the lake bed. In addition there are small fish varieties such as sunbleak, minnow and brook lamprey. Around the lake, grey heron may be seen.The shores of the Titisee are home to two rare types of quillworts, the ...	✗ Atwood Lake. Atwood Lake is a reservoir located in Tuscarawas and Carroll counties in east central Ohio. The lake is formed by Atwood Dam across Indian Fork, a tributary of Conotton Creek. The lake is named for the community of Atwood which was purchased...
UniIR (BLIP_{FF}) ✓inst	 ✗ Wörthersee. ("Coregonus lavaretus"). Etymology.First mentioned as "Werdse" in an 1143 deed, the lake's name originates from the islands in the lake, most notably Maria Wörth, a peninsula since the building of the Glanfurt outlet in 1770...	 ✓ Titisee. The Titisee is a lake in the southern Black Forest in Baden-Württemberg... The lake's outflow, at 840 (m) above sea level, is the River Gutach, which merges with the Haslach stream below Kappel to form the Wutach. The waters of ...	 ✗ Mikorzyńskie Lake. Mikorzyńskie Lake is a lake in central Poland known as Jezioro Mikorzyńskie in Polish. It is located in Konin County in Gmina Ślesin, near the town of Licheń Stary. It is a post-glacial ribbon lake with a strongly developed shoreline...	 ✗ Lake Hallwil. Lake Hallwil (German: "Hallwilersee") is a lake largely in the Canton of Aargau, Switzerland, located at . It is the largest lake in Aargau and lies mostly in the districts of Lenzburg and Kulm on the southern edge of the canton...	 ✗ Głuszyńskie Lake. Głuszyńskie is a lake in Poland near the city of Radziejów in the Kuyavian-Pomeranian Voivodship...
multi-task (BLIP_{FF}) ✗inst	 ✗ Titisee. and burbot on the lake bed. In addition there are small fish varieties such as sunbleak, minnow and brook lamprey. Around the lake, grey heron may be seen. The shores of the Titisee are home to two rare types of quillworts, the...	 ✗ Wdzydze Lake. Wdzydze (Kashubian "Jezoro Wdzydzcz"), also known as the "Kashubian Sea", "Big Water" and "Szerzawa", is a lake in the Tuchola Forest in the Kocierzyna County (Pomeranian Voivodeship), in the region of South Kashubia...	 ✗ Loch Lomond. Powan are one of the commonest fish species in the loch, which has more species of fish than any other loch in Scotland, including lamprey, lampern, brook trout, perch, loach, common roach and flounder...	✗ Mountain Lake (Ontario). Mountain Lake (Ontario) may refer to one of a number of lakes in Ontario, Canada with this precise name or to others with similar names. Mountain Lake (Algoma County)...	✗ Atwood Lake. Atwood Lake is a reservoir located in Tuscarawas and Carroll counties in east central Ohio. The lake is formed by Atwood Dam across Indian Fork, a tributary of Conotton Creek. The lake is named for the community of Atwood...
Zero-shot (BLIP2) ✗inst	 ✗ Lake Annecy. Human-Modified Channel Systems to Lake Annecy, SE France" (2004) Vol 20 (4) Journal of Coastal Research pp...	 ✗ Lake Annecy. Lake Annecy (,) is a perialpine lake in Haute-Savoie in France. It is named after the city of Annecy...	 ✗ Lake Annecy. time the large alpine glaciers melted. It is fed by many small rivers from the surrounding mountains...	 ✗ Monte Pizzocolo. Monte Pizzocolo is a mountain of Lombardy, Italy. It has an elevation of 1,581 metres...	 ✗ Lake Como, Wikimania 2016, MP 126 Lake Como. Wikimania 2016, Esino Lario, Italy...

Figure 16. Visualization of the Top 5 Candidates Retrieved from M-BEIR (5.6M candidates) by Different Models for an Infoseek Query

Task	Dataset	Query Text	Query Image	Retrieved Text	Retrieved Image
1. $q_t \rightarrow c_i$	VisualNews [37]	*Identify the news-related image in line with the described event.* Apple CEO Tim Cook talks about the new iPhones at an event in California 9 September 2015.	-	-	
	MSCOCO [35]	*Find me an everyday image that matches the given caption.* A teddy bear that is on top of a desk.	-	-	
	Fashion200K [18]	*Based on the following fashion description, retrieve the best matching image.* Blue nomad suede side zip mini skirt.	-	-	
2. $q_t \rightarrow c_t$	WebQA [6]	*Retrieve passages from Wikipedia that provide answers to the following question.* Who was the first emperor to visit the the city that is at the head of the Adriatic and the seat of an ancient episcopal see?	-	Augustus was the first of a number of emperors to visit Aquileia, notably during the Pannonian wars in 12–10 BC. It was the birthplace of Tiberius ' son by Julia, in the latter year. The Roman poet Martial praised Aquileia as his hoped for haven and resting place in his old age.	-
3. $q_t \rightarrow (c_i, c_t)$	EDIS [39]	*Find a news image that matches the provided caption.* Bulgarian President Rosen Plevneliev speaks during the 68th session of the UN General Assembly.	-	National SecurityIran leader open to meeting Obama at later date.	
	WebQA [6]	*Find a Wikipedia image that answers this question.* Do both the Hays County Courthouse in San Marcos, Texas and the Ike Wood House at 227 Mitchell Street in San Marcos, Texas have six columns on their front entrance?	-	Hays County Courthouse (2018), San Marcos, TX The Hays County Courthouse in San Marcos, Texas. Listed on the National Register of Historic Places. 227 Mitchell, San Marcos, Texas Ike Wood House at 227 Mitchell Street in San Marcos, Texas.	
4. $q_i \rightarrow c_t$	VisualNews [37]	*Find a caption for the news in the given photo.*		Participants run in front of Jandilla s bulls during the fifth bullrun of the San Fermin Festival in Pamplona northern Spain on July 11.	-
	MSCOCO [35]	*Find an image caption describing the following everyday image.*		A young boy blows on candles on a fire truck cake.	-
	Fashion200K [18]	*Find a product description for the fashion item in the image.*		Blue washed chambray shirt.	-

Figure 17. Examples of datasets in MBEIR. *Query instructions* are written in italic font style.

Task	Dataset	Query Text	Query Image	Retrieved Text	Retrieved Image
5. $q_i \rightarrow c_i$	NIGHTS [15]	*Find a day-to-day image that looks similar to the provided image.*		-	
6. $(q_i, q_t) \rightarrow c_t$	OVEN [19]	*Retrieve a Wikipedia paragraph that provides an answer to the given query about the image.* What is the name of this bridge?		The **Humber Bridge**, near Kingston upon Hull, East Riding of Yorkshire, England, is a 2.22 km (2,430 yd; 7,300 ft; 1.38 mi) single-span road suspension bridge, which opened to traffic on 24 June 1981.	-
	InfoSeek [10]	*Retrieve a Wikipedia paragraph that provides an answer to the given query about the image.* What country does this bridge belong to?		Aqueduct of Segovia. The Aqueduct of Segovia () is a Roman aqueduct in Segovia, **Spain**. It is one of the best-preserved elevated Roman aqueducts and the foremost symbol of Segovia...	-
7. $(q_i, q_t) \rightarrow c_i$	FashionIQ [56]	*Find a fashion image that aligns with the reference image and style note.* Does not have writing in it and has less font and simpler graphics.		-	
	CIRR [40]	*Retrieve a day-to-day image that aligns with the modification instructions of the provided image.* White fuzzy animal sits by the food bowl.		-	
8. $(q_i, q_t) \rightarrow (c_i, c_t)$	OVEN [19]	*Retrieve a Wikipedia image-description pair that provides evidence for the question of this image.* What is the model of this aircraft?		Identify the news-related image that aligns with the described event. The **British Aerospace 146** (also BAe 146) is a short-haul and regional airliner that was manufactured in the United Kingdom by British Aerospace, later part of BAE Systems.	
	InfoSeek [10]	*Retrieve a Wikipedia image-description pair that provides evidence for the question of this image.* What is the country of origin of this animal?		The Old English Sheepdog is a large breed of dog that emerged in **England** from early types of herding dog.	

Figure 18. Additional examples of datasets in MBEIR. *Query instructions* are written in italic font style.

Task	Dataset	Query Instruction
1. $q_t \rightarrow c_i$	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 you 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.
2. $q_t \rightarrow c_t$	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? RI'm looking for a Wikipedia snippet that answers this question.
3. $q_t \rightarrow (c_i, c_t)$	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.
4. $q_i \rightarrow c_t$	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.

Table 9. M-BEIR query instructions.

Task	Dataset	Query Instruction
5. $q_i \rightarrow c_t$	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.
6. $(q_i, q_t) \rightarrow c_t$	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.
7. $(q_i, q_t) \rightarrow c_i$	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.
8. $(q_i, q_t) \rightarrow (c_i, c_t)$	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 10. M-BEIR query instructions.

			Zero-shot				**Multi-task (✗ instruction)**					**UniIR (✓ instruction)**				
Task	Dataset	Metric	CLIP	SigLIP	BLIP	BLIP2	CLIP_{SF}	CLIP_{FF}	BLIP_{SF}	BLIP_{FF}	$\text{BLIP}_{\text{FF, 384}}$	CLIP_{SF}	CLIP_{FF}	BLIP_{SF}	BLIP_{FF}	$\text{BLIP}_{\text{FF, 384}}$
1. $q_t \to c_i$	VisualNews	R@1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	22.3	12.8	8.4	9.9	11.8
		R@5	0.0	0.0	0.0	0.0	12.7	8.8	5.0	8.3	10.5	42.6	28.8	20.9	23.0	26.5
		R@10	0.0	0.0	0.0	0.0	22.3	15.8	9.4	14.5	17.2	51.9	37.1	27.4	29.9	34.0
	MSCOCO	R@1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	45.8	47.5	23.3	37.4	20.4
		R@5	0.0	0.0	0.0	0.0	27.3	24.6	22.9	27.7	35.2	77.9	74.7	71.6	75.6	75.7
		R@10	0.0	0.0	0.0	0.0	47.3	42.8	41.0	45.8	53.9	86.0	83.5	81.1	84.3	84.6
	Fashion200K	R@10	0.0	0.0	0.0	0.0	5.9	5.9	5.7	9.0	13.1	17.8	15.5	24.3	25.4	26.7
		R@20	0.0	0.0	0.0	0.0	12.1	10.9	12.4	17.4	22.2	24.5	21.6	32.8	35.4	34.3
		R@50	0.0	0.0	0.0	0.0	25.6	21.0	26.7	30.3	36.4	36.2	32.8	45.4	48.3	48.1
2. $q_t \to c_t$	WebQA	R@1	17.4	17.0	21.2	20.7	56.4	42.4	48.3	49.0	48.5	58.3	51.6	51.6	51.6	52.8
		R@5	32.1	34.0	38.1	35.2	82.3	67.9	74.4	76.1	76.9	84.7	78.4	78.9	79.5	79.2
		R@10	38.9	41.5	45.7	41.5	87.5	75.6	81.2	82.5	83.2	89.5	85.4	84.9	85.2	85.5
3. $q_t \to (c_i, c_t)$	EDIS	R@1	1.4	0.1	0.0	0.0	15.2	12.8	13.5	14.0	15.7	30.1	24.5	23.3	25.2	25.7
		R@5	6.7	1.1	0.0	0.0	41.1	38.3	33.6	36.0	38.5	59.4	50.0	47.2	50.3	51.4
		R@10	9.8	2.1	0.0	0.0	54.4	51.2	43.4	45.9	49.8	70.4	60.8	56.8	60.9	63.2
	WebQA	R@1	1.7	0.4	0.0	0.0	38.9	34.6	43.2	43.7	46.4	50.1	48.8	48.5	50.9	51.5
		R@5	5.5	2.1	0.0	0.0	68.2	62.5	73.2	74.7	75.2	78.8	75.3	76.8	79.7	79.4
		R@10	7.8	3.6	0.0	0.0	78.8	72.8	82.9	83.0	83.5	86.5	83.5	85.3	87.1	87.1
4. $q_i \to c_t$	VisualNews	R@1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	22.6	12.6	5.2	7.4	9.2
		R@5	0.0	0.0	0.0	0.0	12.1	8.2	4.8	4.9	6.0	42.8	28.6	19.4	21.1	22.9
		R@10	0.0	0.0	0.0	0.0	22.4	15.4	9.3	9.6	11.4	52.2	37.0	26.9	28.7	30.4
	MSCOCO	R@1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	73.8	66.9	39.7	53.4	62.4
		R@5	0.0	0.0	0.0	0.0	84.6	80.8	74.9	76.9	81.4	92.3	89.0	88.2	88.8	90.1
		R@10	0.0	0.0	0.0	0.0	92.3	90.2	85.9	87.5	91.1	96.0	94.5	94.1	94.8	95.7
	Fashion200K	R@10	0.0	0.0	0.0	0.0	1.2	1.3	2.6	3.6	4.0	17.9	13.7	24.3	27.6	28.4
		R@20	0.0	0.0	0.0	0.0	2.9	3.0	5.9	7.5	8.5	25.7	20.4	33.5	37.4	38.3
		R@50	0.0	0.0	0.0	0.0	7.3	6.5	13.9	16.6	17.9	38.1	32.6	46.2	51.6	52.3
5. $q_i \to c_i$	NIGHTS	R@1	6.5	7.5	6.7	6.3	8.3	7.5	9.0	8.5	9.1	8.3	7.3	8.5	8.2	8.8
		R@5	25.3	28.7	25.1	24.0	31.0	30.8	32.9	31.3	32.5	32.0	31.9	33.4	33.0	33.7
		R@10	42.2	46.0	40.1	38.1	52.4	52.9	52.8	52.8	55.5	53.7	52.1	53.8	53.7	54.4
6. $(q_i, q_t) \to c_t$	OVEN	R@1	0.0	0.0	0.0	0.0	22.2	19.5	20.0	24.6	28.4	26.1	22.4	21.3	25.6	27.4
		R@5	0.0	0.0	0.0	0.0	36.8	31.6	33.2	37.7	39.2	39.2	34.7	35.2	38.7	40.7
		R@10	0.0	0.0	0.0	0.0	43.2	37.6	39.3	43.7	44.3	45.1	40.5	41.6	44.8	46.4
	InfoSeek	R@1	0.0	0.0	0.0	0.0	8.4	7.0	4.6	7.6	8.7	11.4	8.5	7.5	9.2	9.5
		R@5	0.0	0.0	0.0	0.0	18.3	15.4	11.9	17.8	17.1	24.0	17.5	16.7	19.7	19.2
		R@10	0.0	0.0	0.0	0.0	25.0	21.8	16.9	23.5	22.6	31.2	23.2	22.4	25.9	25.2
7. $(q_i, q_t) \to c_i$	FashionIQ	R@10	4.4	4.8	2.2	3.9	22.8	19.7	26.1	28.1	29.0	24.3	20.5	26.2	28.5	29.8
		R@20	6.7	6.5	3.7	6.2	30.7	26.4	33.6	35.9	36.2	32.5	27.8	34.1	36.2	36.5
		R@50	11.6	9.7	6.1	10.1	42.1	36.7	45.6	47.6	47.1	43.2	39.5	46.1	46.9	47.1
	CIRR	R@1	0.7	0.5	0.8	0.9	4.1	10.9	11.5	20.7	23.5	8.2	13.1	11.9	24.4	25.0
		R@5	5.4	7.1	7.4	6.2	32.0	32.7	36.7	45.1	47.4	43.9	40.9	43.0	51.4	51.1
		R@10	8.2	10.8	11.3	9.0	43.2	42.9	47.7	55.8	58.2	56.8	52.7	54.8	61.8	61.4
8. $(q_i, q_t) \to (c_i, c_t)$	OVEN	R@1	14.9	18.1	4.4	7.8	44.1	37.0	36.4	37.6	40.5	45.9	42.2	36.0	44.5	46.9
		R@5	24.5	27.2	10.1	13.8	58.7	50.1	51.0	51.6	53.1	60.2	55.8	51.8	57.8	59.5
		R@10	29.5	32.0	13.5	17.4	63.8	55.7	57.0	57.4	58.2	65.4	61.2	58.6	62.8	64.4
	InfoSeek	R@1	7.0	8.7	1.7	3.4	22.4	16.0	11.4	12.3	12.6	25.7	18.6	11.4	13.7	16.6
		R@5	22.1	24.3	7.9	11.4	42.3	31.5	23.0	25.4	25.2	44.6	36.8	25.4	27.7	31.1
		R@10	29.7	31.4	11.9	16.0	50.9	39.5	30.4	32.8	32.2	53.8	45.2	32.9	35.9	38.8
-	Average	R@1	3.4	3.6	2.3	2.7	15.6	13.4	14.5	16.2	17.5	30.5	26.7	23.2	27.7	28.3
		R@5	8.0	8.2	5.8	6.1	37.1	32.7	33.1	35.9	37.8	50.3	44.5	44.3	47.2	48.1
		R@10	11.1	11.1	8.0	8.3	47.4	42.4	42.7	45.6	47.7	59.8	53.8	53.6	56.4	57.4

Table 11. Benchmarking information retrieval recall@1/5/10 on M-BEIR. For Fashion200K and FashionIQ, we report recall@10/20/50 following original work.

			Multi-task (✗ instruction)				UniIR (✓ instruction)			
			CLIP_{SF}		CLIP_{FF}		CLIP_{SF}		CLIP_{FF}	
	Dataset	Metric	$\text{M-BEIR}_{\text{local}}$	M-BEIR	$\text{M-BEIR}_{\text{local}}$	M-BEIR	$\text{M-BEIR}_{\text{local}}$	M-BEIR	$\text{M-BEIR}_{\text{local}}$	M-BEIR
1. $q_t \rightarrow c_i$	VisualNews	R@1	10.4	0.0	6.4	0.0	11.8	11.5	7.0	6.9
		R@5	24.1	2.9	16.6	3.5	26.3	26.1	18.1	18.1
		R@10	31.0	6.0	22.6	7.5	34.1	34.0	25.1	25.0
	MSCOCO	R@1	41.8	0.0	38.0	0.0	43.8	36.1	39.5	36.3
		R@5	70.3	11.6	67.0	12.8	72.1	67.2	68.9	64.6
		R@10	80.1	27.0	77.8	28.0	82.0	77.2	79.6	75.2
	Fashion200K	R@10	10.2	3.3	9.6	3.9	9.7	9.6	9.9	9.9
		R@20	15.7	7.2	14.9	7.5	14.4	14.4	13.4	13.4
		R@50	26.4	16.9	23.9	16.9	22.8	22.8	22.8	22.8
2. $q_t \rightarrow c_t$	WebQA	R@1	55.1	53.7	49.0	47.8	54.0	54.0	51.3	51.3
		R@5	82.0	80.5	74.1	72.9	81.1	81.1	79.1	79.1
		R@10	87.9	86.0	81.7	80.3	87.8	87.8	85.6	85.6
3. $q_t \rightarrow (c_i, c_t)$	EDIS	R@1	22.7	12.5	22.7	11.4	25.7	25.8	23.1	23.1
		R@5	46.8	36.4	46.8	34.6	50.9	50.8	46.3	46.2
		R@10	57.3	47.6	56.7	45.5	60.9	60.9	56.2	56.2
	WebQA	R@1	46.4	37.5	43.2	34.7	46.8	46.9	46.4	46.8
		R@5	74.3	67.0	71.5	63.4	74.5	74.7	74.4	74.4
		R@10	83.6	77.3	81.0	73.9	84.0	83.9	83.8	83.8
4. $q_i \rightarrow c_t$	VisualNews	R@1	10.4	0.0	6.5	0.0	12.5	12.2	7.4	7.4
		R@5	23.8	4.0	16.2	4.5	27.1	26.8	18.4	18.4
		R@10	31.0	8.1	22.2	8.6	34.9	34.6	25.0	25.0
	MSCOCO	R@1	55.6	0.0	50.3	0.0	60.8	60.7	55.5	55.5
		R@5	80.7	67.5	77.9	69.0	84.0	84.0	81.8	81.8
		R@10	88.8	79.6	86.8	81.0	90.6	90.6	89.5	89.5
	Fashion200K	R@10	11.6	0.7	9.7	0.8	10.3	10.2	9.1	9.1
		R@20	16.7	1.5	14.5	2.2	15.3	15.3	13.8	13.8
		R@50	26.8	4.0	24.5	5.4	25.1	25.0	22.8	22.8
5. $q_i \rightarrow c_i$	NIGHTS	R@1	8.1	8.1	7.3	7.3	8.0	8.0	8.1	8.1
		R@5	29.6	29.2	27.1	26.9	31.9	31.9	29.1	29.0
		R@10	49.2	48.4	46.1	45.3	48.8	48.7	48.3	48.2
6. $(q_i, q_t) \rightarrow c_t$	OVEN	R@1	16.9	15.0	15.9	13.9	16.4	16.4	15.2	15.5
		R@5	33.9	26.8	31.9	24.2	32.6	28.2	31.6	26.6
		R@10	41.6	32.6	39.4	29.6	40.1	34.0	39.2	32.4
	InfoSeek	R@1	7.0	4.4	7.1	3.8	7.0	5.4	6.6	5.3
		R@5	16.8	10.6	16.3	9.7	16.9	13.3	15.6	12.9
		R@10	22.2	14.5	21.9	13.6	22.7	18.4	20.5	17.1
7. $(q_i, q_t) \rightarrow c_i$	FashionIQ	R@10	16.3	15.7	13.2	12.8	17.3	16.5	14.1	14.0
		R@20	22.5	21.7	19.0	18.6	23.4	22.5	20.0	19.8
		R@50	33.0	32.2	28.7	27.9	34.3	32.8	30.5	30.1
	CIRR	R@1	2.4	2.2	7.4	6.0	2.6	2.6	9.7	9.6
		R@5	34.3	23.2	31.2	23.0	35.6	33.2	35.1	34.1
		R@10	45.4	32.3	42.4	32.0	48.0	45.9	48.5	47.5
8. $(q_i, q_t) \rightarrow (c_i, c_t)$	OVEN	R@1	36.1	29.1	31.9	27.3	34.5	32.6	31.4	30.6
		R@5	55.5	42.1	50.7	39.8	54.5	46.5	50.9	43.7
		R@10	62.7	47.8	58.2	45.3	61.8	52.5	58.2	49.6
	InfoSeek	R@1	16.3	11.6	13.1	10.0	16.8	14.8	13.7	11.8
		R@5	32.6	24.3	27.0	21.2	33.1	29.1	28.8	25.6
		R@10	41.0	31.2	34.4	26.9	41.2	36.9	36.8	33.0
-	Average	R@1	25.3	13.4	23.0	12.5	26.2	25.1	24.2	23.7
		R@5	46.5	32.8	42.6	31.2	47.7	45.6	44.5	42.7
		R@10	47.5	34.9	44.0	33.5	48.4	46.4	45.6	43.8

Table 12. Benchmarking information retrieval recall@1/5/10 on M-BEIR for CLIP Base models. For Fashion200K and FashionIQ, we report recall@10/20/50 following the original work.

			Multi-task (✗ instruction)				UniIR (✓ instruction)			
			BLIP_{SF}		BLIP_{FF}		BLIP_{SF}		BLIP_{FF}	
	Dataset	Metric	$\text{M-BEIR}_{\text{local}}$	M-BEIR	$\text{M-BEIR}_{\text{local}}$	M-BEIR	$\text{M-BEIR}_{\text{local}}$	M-BEIR	$\text{M-BEIR}_{\text{local}}$	M-BEIR
1. $q_t \rightarrow c_i$	VisualNews	R@1	7.1	0.0	7.8	0.0	6.9	5.5	7.5	6.9
		R@5	16.9	2.3	18.2	3.4	17.0	15.9	18.0	17.6
		R@10	22.9	5.0	24.3	6.9	22.8	21.9	24.3	23.8
	MSCOCO	R@1	47.3	0.0	48.1	0.0	47.8	11.3	49.5	23.8
		R@5	74.6	15.4	75.7	10.5	75.6	65.1	76.6	70.2
		R@10	83.4	32.9	84.5	24.4	84.5	76.5	85.1	80.2
	Fashion200K	R@10	22.1	3.5	22.3	0.9	21.6	21.2	22.8	22.1
		R@20	29.7	7.5	29.6	3.0	28.2	27.8	30.6	30.3
		R@50	42.5	18.2	41.8	8.7	40.0	39.7	41.5	41.1
2. $q_t \rightarrow c_t$	WebQA	R@1	50.1	47.9	49.2	46.9	51.4	51.2	50.1	49.7
		R@5	77.0	74.2	76.2	74.6	76.7	76.7	76.4	76.4
		R@10	84.0	82.0	83.4	81.9	83.3	83.2	82.7	82.7
3. $q_t \rightarrow (c_i, c_t)$	EDIS	R@1	20.7	12.0	23.8	12.2	22.0	21.2	23.5	22.6
		R@5	40.2	30.0	47.5	30.4	44.5	44.1	46.2	45.9
		R@10	50.1	38.7	57.7	38.9	53.9	53.7	56.7	56.4
	WebQA	R@1	49.0	45.0	49.1	45.7	49.9	49.6	49.5	49.1
		R@5	77.3	73.5	78.1	75.0	77.5	77.2	78.1	77.9
		R@10	86.6	83.4	85.9	82.6	86.4	86.2	85.6	85.5
4. $q_i \rightarrow c_t$	VisualNews	R@1	7.1	0.0	7.4	0.0	6.6	3.7	7.2	5.0
		R@5	17.1	3.2	17.1	2.3	16.8	14.8	17.1	15.8
		R@10	23.2	6.7	22.8	5.0	22.6	21.4	23.6	22.5
	MSCOCO	R@1	58.3	0.0	63.7	0.0	62.4	39.2	64.6	51.7
		R@5	83.0	71.6	87.0	71.7	85.9	84.5	88.1	86.9
		R@10	90.3	83.3	93.1	83.9	92.0	91.5	93.5	93.1
	Fashion200K	R@10	22.4	1.6	23.6	0.9	20.9	18.2	23.8	22.4
		R@20	30.4	4.0	32.0	2.4	29.7	27.7	32.7	31.5
		R@50	43.8	10.2	44.7	6.8	42.3	40.6	45.6	45.1
5. $q_i \rightarrow c_i$	NIGHTS	R@1	8.2	8.2	8.0	8.0	7.7	7.7	7.7	7.7
		R@5	30.0	29.7	31.8	31.6	30.7	30.6	30.3	30.3
		R@10	49.8	49.1	51.0	50.5	51.2	51.2	49.5	49.5
6. $(q_i, q_t) \rightarrow c_t$	OVEN	R@1	17.6	16.9	19.6	20.4	13.9	16.0	17.9	20.8
		R@5	33.2	29.2	36.7	32.8	28.9	28.2	34.5	33.1
		R@10	40.7	35.2	44.4	38.9	36.1	34.1	42.1	39.0
	InfoSeek	R@1	7.6	3.7	8.6	5.6	6.3	5.0	7.9	6.9
		R@5	17.3	9.5	21.2	13.6	16.5	13.3	18.9	16.5
		R@10	23.6	14.0	28.1	19.3	23.3	19.6	25.4	22.5
7. $(q_i, q_t) \rightarrow c_i$	FashionIQ	R@10	22.5	22.1	25.4	24.9	20.8	20.1	23.7	23.0
		R@20	29.9	29.2	33.2	32.5	27.5	26.6	31.0	29.8
		R@50	41.2	40.3	44.8	43.9	38.2	36.6	42.1	40.5
	CIRR	R@1	11.0	9.1	20.3	18.7	13.1	12.9	20.2	19.4
		R@5	39.0	31.8	45.1	42.0	42.2	40.7	46.4	45.1
		R@10	50.3	42.7	56.0	52.2	54.4	52.3	57.5	55.8
8. $(q_i, q_t) \rightarrow (c_i, c_t)$	OVEN	R@1	28.3	32.5	31.1	33.2	28.3	33.0	29.1	37.1
		R@5	47.0	46.8	49.9	47.5	47.3	48.2	48.2	50.6
		R@10	54.6	52.5	57.5	53.1	54.7	54.2	55.6	55.9
	InfoSeek	R@1	11.2	9.4	13.7	10.9	12.2	10.8	13.0	11.8
		R@5	25.5	20.7	29.1	22.7	26.4	23.6	27.0	23.5
		R@10	33.6	27.9	37.3	29.4	34.2	30.7	35.0	30.4
-	Average	R@1	24.9	14.2	26.9	15.5	25.3	20.6	26.8	24.1
		R@5	44.5	33.7	47.2	35.2	45.1	43.3	46.6	45.4
		R@10	47.5	36.3	49.8	37.1	47.7	46.0	49.2	47.8

Table 13. Benchmarking information retrieval recall@1/5/10 on M-BEIR for BLIP Base models. For Fashion200K and FashionIQ, we report recall@10/20/50 following the original work.