UniIR: Training and Benchmarking Universal Multimodal Information Retrievers

Cong Wei[1], Yang Chen[2], Haonan Chen[1], Hexiang Hu[4], Ge Zhang[1], Jie Fu[3], Alan Ritter[2], and Wenhu Chen[1]

[1] University of Waterloo
[2] Georgia Institute of Technology
[3] Hong Kong University of Science and Technology
[4] Google DeepMind
{cong.wei,wenhuchen}@uwaterloo.ca, {yang.chen,alan.ritter}@cc.gatech.edu

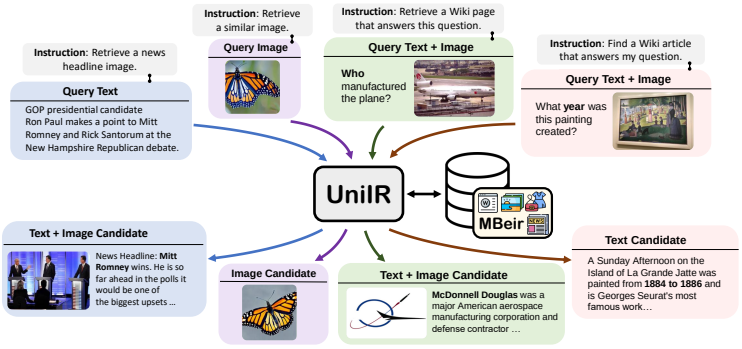


Fig. 1: We build a universal multimodal information retriever UniIR through instruction tuning. UniIR is capable of accepting any form of query and instruction to retrieve information in any modality.

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.

Keywords: Cross-modal / multi-modal retrieval · Image retrieval

1 Introduction

Information retrieval (IR) is a pivotal task that involves sourcing relevant information from vast data collections to meet specific user requirements [46]. This process has become increasingly important with the advent of generative AI [4, 9, 45, 60], as it not only enables precise attribution but also mitigates the risk of inaccuracies and fabrications in generated content [2, 44]. 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 [30] considers retrieving Flickr images via text caption, while EDIS [34] 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 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 [43] 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 [43] and BLIP [28] 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 dataset-specific 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

Task (query → candidate)	**Dataset**	**Instruction** (shown 1 out of 4)	**Domain**	**Train**	**Dev**	**Test**	**Pool**
1. $q_t \rightarrow c_i$	VisualNews [32]	Identify news-related image match with the description	News	99K	20K	20K	542K
	MSCOCO [30]	Find an everyday image match with caption	Misc.	100K	24.8K	24.8K	5K
	Fashion200K [19]	Based on fashion description, retrieve matched image	Fashion	15K	1.7K	1.7K	201K
2. $q_t \rightarrow c_t$	WebQA [7]	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 [34]	Find a news image matching with the caption	News	26K	3.2K	3.2K	1M
	WebQA [7]	Find a Wiki image that answer the question	Wiki	17K	1.7K	2.5K	403K
4. $q_i \rightarrow c_t$	VisualNews [32]	Provide a news-related caption for the displayed image	News	100K	20K	20K	537K
	MSCOCO [30]	Find a caption describe the an image	Misc.	113K	5K	5K	25K
	Fashion200K [19]	Find a description for the fashion item in the image	Fashion	15K	4.8K	4.8K	61K
5. $q_i \rightarrow c_i$	NIGHTS [16]	Find an image that is identical to the given image	Misc.	16K	2K	2K	40K
6. $(q_i, q_t) \rightarrow c_t$	OVEN [21]	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 [36]	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 [21]	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.

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.

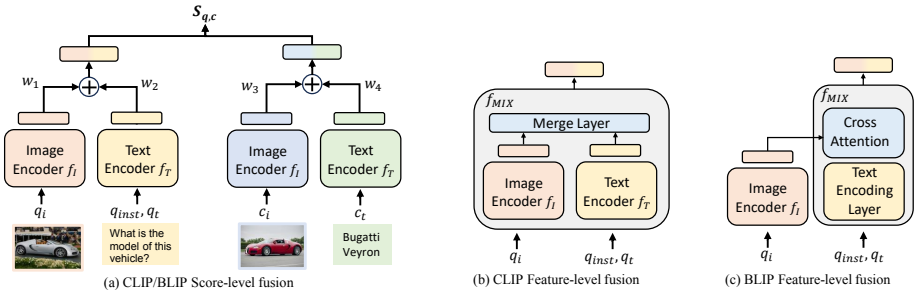
2.2 UniIR Model

In this section, we present the UniIR model, our unified multimodal information retrieval system. The UniIR model is capable of handling multiple retrieval tasks at once. We build the UniIR model based on pre-trained vision-language models CLIP [43] and BLIP [28]. These two models are only applicable for image and text cross-modality retrieval tasks. To adapt them as UniIR models, which can handle multi-modality retrieval tasks such as image-to-image-text pair retrieval, we proposed two simple yet effective multi-modal fusion mechanisms, namely score-level fusion and feature-level fusion [21, 34, 35].

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 unimodal 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.



(a) CLIP/BLIP Score-level fusion (b) CLIP Feature-level fusion (c) BLIP Feature-level fusion

Fig. 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.

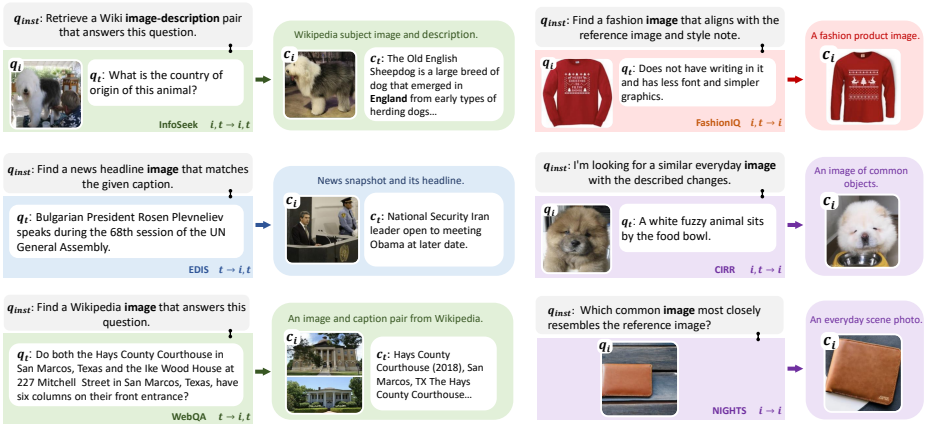


Fig. 3: Examples of six query instances in the M-BEIR dataset. Each example query instance includes a query $\mathbf{q}$, a human-annotated natural language instruction q_{inst}, and a positive(relevant) candidate $\mathbf{c}^+$.

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), in BLIP feature-level fusion (BLIP$_{\text{FF}}$), we discard the BLIP's text encoder and leverage the image-grounded text encoder instead. The process begins with the extraction of image embeddings through the vision encoder f_I. These embeddings are then integrated with text embeddings through the cross-attention layers of the image-grounded text encoder, labelled as f_{MIX}. For the CLIP feature-level fusion (CLIP$_{\text{FF}}$), 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 Transformer, forming a mixed-modality encoder f_{MIX}. In both CLIP$_{\text{FF}}$ and BLIP$_{\text{FF}}$, the output from f_{MIX} is a comprehensive feature vector that combines information from both image and text modalities. The final repre-

sentations 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 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 [21], EDIS [34], CIRR [36] and FashionIQ [56]), image-caption datasets

(MS-COCO [30], Fashion200K [19], VisualNews [32]), image-similarity measurement dataset (NIGHTS [16]), along with retrieval-based VQA datasets (InfoSeek [10], WebQA [7]). 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 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 [13, 55]. To design instructions for multimodal retrieval tasks, we took inspiration from the instruction schema in TART [3]. 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 [19], 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.

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 [43]/BLIP [28] 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 can be found in the Appendix.

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

		SoTA Zero-Shot							UniIR			
Task	**Dataset**	NExt-GPT	GPT-4V	CoDi	CLIP	SigLIP	BLIP	BLIP2	CLIP_{SF}	CLIP_{FF}	BLIP_{SF}	BLIP_{FF}
1. $q_t \rightarrow c_i$	VisualNews	0.0	0.0	11.4	0.0	0.0	0.0	0.0	**42.6**	28.8	20.9	23.0
	MSCOCO	6.0	0.0	1.0	0.0	0.0	0.0	0.0	**77.9**	74.7	71.6	75.6
	Fashion200K	0.0	5.0	1.0	0.0	0.0	0.0	0.0	17.8	15.5	24.3	**25.4**
2. $q_t \rightarrow c_t$	WebQA	20.0	56.1	0.0	32.1	34.1	38.1	35.2	**84.7**	78.4	78.9	79.5
3. $q_t \rightarrow (c_i, c_t)$	EDIS	0.0	6.0	0.0	6.7	1.1	0.0	0.0	**59.4**	50.0	47.2	50.3
	WebQA	0.0	12.0	0.0	5.5	2.2	0.0	0.0	78.8	75.3	76.8	**79.7**
4. $q_i \rightarrow c_t$	VisualNews	0.0	4.0	0.9	0.0	0.0	0.0	0.0	**42.8**	28.6	19.4	21.1
	MSCOCO	16.2	72.5	5.7	0.0	0.0	0.0	0.0	**92.3**	89.0	88.2	88.8
	Fashion200K	0.0	14.2	0.0	0.0	0.0	0.0	0.0	17.9	13.7	24.3	**27.6**
5. $q_i \rightarrow c_t$	NIGHTS	0.0	30.1	0.9	25.3	28.7	25.1	24.0	32.0	31.9	**33.4**	33.0
6. $(q_i, q_t) \rightarrow c_t$	OVEN	0.0	25.0	0.0	0.0	0.0	0.0	0.0	**39.2**	34.7	35.2	38.7
	InfoSeek	0.0	12.3	0.0	0.0	0.0	0.0	0.0	**24.0**	17.5	16.7	19.7
7. $(q_i, q_t) \rightarrow c_i$	FashionIQ	0.0	0.0	3.8	6.7	6.5	3.7	6.3	24.3	20.5	26.2	**28.5**
	CIRR	0.0	0.0	0.0	5.4	7.1	7.4	6.2	43.9	40.9	43.0	**51.4**
8. $(q_i, q_t) \rightarrow (c_i, c_t)$	OVEN	0.0	12.1	0.0	24.5	27.2	10.1	13.8	**60.2**	55.8	51.8	57.8
	InfoSeek	0.0	18.2	0.0	22.1	24.3	7.9	11.4	**44.6**	36.8	25.4	27.7
-	Average	2.6	16.7	1.6	8.0	8.2	5.8	6.1	**48.9**	43.3	42.7	45.5

Table 2: Benchmarking SoTA zero-shot baselines and UniIR on retrieving from a heterogeneous candidate pool of M-BEIR (5.6M candidates) with Recall@5 (except using Recall@10 for Fashion200K, FashionIQ). **Bold**: top-1 performance. Underline: top-2.

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 dataset-specific pool (with homogeneous candidates) provided by the original dataset, which enables a fair comparison with existing SoTA retrievers. We name this dataset-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 candidates in the pool. We then utilize FAISS [24], to index and retrieve candidates.

4.1 Baselines

Zero-shot SoTA Retriever. We utilize pre-trained vision-language models such as CLIP (L-14)[43], SigLIP (L)[62], BLIP (L) [28], and BLIP2 [27] as our baseline retrievers. In addition, we adapt pre-trained multi-modal multi-task generative models such as CoDi [50], NExt-GPT [57], and GPT-4V(DALLE3) [1, 59] as our baselines. These generative models are general-purpose and can leverage the reasoning ability of LLMs, making them a strong baseline for our experiments. To adapt them for retrieval tasks, we prompt them to produce a guessed candidate based on the query and instruction (e.g., generate an image based on text description), then assess its similarity to M-BEIR candidates using CLIP embedding space. The caveat is that baseline models may not fully understand the intent of the retrieval task, thus, they are expected to achieve lower performance in the standard setting (1) with a heterogeneous candidate pool. We do not evaluate any single-stream models or modules such as UNITER [11] or prompt

GPT-4V to generate a similarity score for every query-candidate pair. These methods are inefficient and lead to long execution time for large-scale retrieval.

Dataset-specific/Multi-task Fine-tuned Baselines. We fine-tune CLIP and BLIP on each dataset as our dataset-specific (DS) baseline retrievers. We also fine-tune CLIP and BLIP jointly on all M-BEIR training data without incorporating instructions as our multi-task (MT) 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 [15]. 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 due to memory constrain. 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

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

Table 3: Benchmarking UniIR and multi-task tuning baselines on retrieving from a heterogeneous candidate pool of M-BEIR (5.6M candidates) with Recall@5 (except using Recall@10 for Fashion200K, FashionIQ). Multi-task baselines are tuned without incorporating instructions. Δ: UniIR - Multi-task. **Bold**: top-1 performance. Underline: top-2.

We report the main results on M-BEIR in Tab. 2, where models retrieve candidates from the 5.6M pool. We show that zero-shot models struggle to retrieve queried information from such a heterogeneous pool. We demonstrate

Dataset	Domain	Task	Query Instruction q_{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...	✗ 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 ...	✗ President Obama walks with Germany s Chancellor Angela Merkel in St Petersburg Russia on Sept 6.

Fig. 4: Visualization of top 5 retrieved candidates from M-BEIR with 3 models on EDIS. Zero-shot and multi-task training models mostly retrieve the wrong modality (text-only). UniIR retrieves candidates accurately with the right modality (image, text).

instruction-tuning as a crucial component in Table 3. Furthermore, we also conduct experiments to understand the zero-shot generalization of UniIR, where we train UniIR on a subset of datasets and evaluate on the held-out dataset set, which makes UniIR fairly comparable with other zero-shot retrievers.

Zero-shot retrievers cannot comprehend retrieval intention. We benchmark four open-sourced cross-modal embedding models(CLIP, SigLIP, BLIP and BLIP2) and three multi-modal generative models(NExt-GPT, GPT-4V(DALLE3) and CoDi) in Table 2. We found that the recall values on most tasks are near zero. These pre-trained models struggle to comprehend the task intention. For example, in the text-to-image retrieval task on MSCOCO, all the cross-modal embedding 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. In Figure 4, we present examples where BLIP2 retrieves distracting candidates from the wrong modality for an EDIS query. We also observed that multi-modal generative models often generate inaccurate guess candidates, making it difficult to identify the ground-truth candidate from the global pool, especially when the target candidate contains an image.

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 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.

	Zero-Shot	Multi-Task	UniIR	Zero-Shot	Multi-Task	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 4: Experiments of instruction-tuning. Retrieve from the M-BEIR (Recall@5). Δ: UniIR - Multi-task (Multi.). Results are obtained by averaging Tab. 3 results across datasets.

	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 5: Error analysis on M-BEIR. UniIR has exhibited a superior ability to follow instructions as compared to the multi-task baselines which were trained without instructions.

UniIR can precisely follow instructions. To further demonstrate the advantages of UniIR over Multi-task finetuning 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 5. 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.

UniIR can generalize to unseen retrieval datasets. 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 datasets compared to their multi-task counterparts without instructions.

		SoTA Zero-Shot				**DS**	**MT**	**UniIR**	**DS**	**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: Experiments on retrieving from a single dataset candidate pool ($\text{M-BEIR}_{\text{local}}$ setting). We report Recall@5 results of zero-shot retrieval, dataset-specific (DS) fine-tuning, Multi-task tuning (MT) baselines and UniIR on $\text{M-BEIR}_{\text{local}}$ except for Fashion200K and FashionIQ where we report Recall@10. Δ_s: absolute difference to dataset-specific fine-tuning. We have omitted CoDi, NExt-GPT, and GPT-4V from this table due to page constraints and they perform worse than the other zero-shot models. The full table can be found in the Appendix.

4.3 Comparison with Existing Methods on Retrieving from Dataset-specific Pool

To compare UniIR with existing retrievers, we also evaluate the homogeneous setting where the retriever only needs to retrieve from the dataset-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 dataset-specific pools $\text{M-BEIR}_{\text{local}}$ in comparison to baseline models. In this section, we focus on UniIR (CLIP_{SF}) and UniIR (BLIP_{FF}), as they have better performance than the other fusion mechanism counterparts as shown in 3.

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

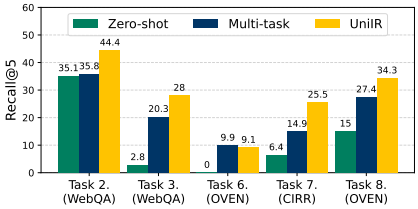


Fig. 5: Held-out dataset generalization experiments on M-BEIR: we train a Multi-task and a UniIR model on 7 held-in datasets and test on 3 held-out datasets (WebQA, OVEN, CIRR) from the M-BEIR. Results are averaged over CLIP_{SF} and BLIP_{FF}.

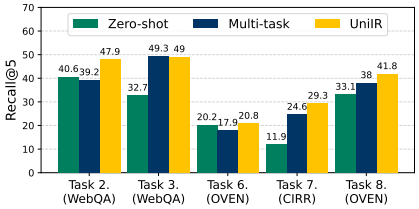


Fig. 6: Held-out dataset generalization experiments on $\text{M-BEIR}_{\text{local}}$: we train a Multi-task and a UniIR model on 7 held-in datasets and test on 3 held-out datasets (WebQA, OVEN, CIRR). Results are averaged over CLIP_{SF} and BLIP_{FF}.

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 Dataset-specific Tuning. Table 6 demonstrates the advantages of multi-task instruction-tuning in the UniIR framework over dataset-specific fine-tuning. Our findings indicate that UniIR (BLIP_{FF}) greatly outperforms its dataset-specific 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 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 UniIR models exhibit superior generalization ability on unseen 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 [25] and Flickr30k [42] have become standard benchmarks for evaluating the progress of pre-trained vision-language models such as ALIGN [23], VILT [26], ALBEF [29], MURAL [22], and

ImageBind [18]. However, fine-grained image retrieval often hinges on the ability to articulate intents through text, presenting challenges in multimodal queries [8] such as ReMuQ [37]. 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 M-BEIR will accelerate progress toward more general multimodal information retrieval models.

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) [12, 39, 53]. FLAN [13, 55] have demonstrated capabilities to generalize to unseen natural language tasks or instructions [38, 40] Recently, visual instruction tuning [14, 33] has been explored in vision-language tasks [58]. On image diffusion models, InstructPix2Pix [5], MagicBrush [63], and Instruct-Imagen [20] 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 [31], as well as embedding models like OneEmbedder [47] and TART [3], remain text-only. In contrast, UniIR demonstrates promising cross-dataset generalization in multimodal retrieval.

Multimodal multitask generative model. Recent advances in multimodal generative models, such as Gemini [51], have shown significant progress in building foundational models capable of reading and generating multimodal inputs and outputs. These models have demonstrated their capabilities to follow instructions for generating text [17], images [57], and audio [49, 50] using adapted diffusion models. Another line of work discretize multimodal data into tokens and process them in an autoregressive fashion [6, 41, 48, 61]. Although these models are not designed for retrieval tasks, we leverage the instructional-following capabilities of these models to generate text/image directly, in conjunction with a CLIP model for single modality retrieval (Table 2). While the current performance on retrieval are outperformed, they show potential for future adaption to retrieval.

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.

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.

Bibliography

[1] Achiam, J., Adler, S., Agarwal, S., Ahmad, L., Akkaya, I., Aleman, F.L., Almeida, D., Altenschmidt, J., Altman, S., Anadkat, S., et al.: Gpt-4 technical report. arXiv preprint arXiv:2303.08774 (2023) 8

[2] Asai, A., Min, S., Zhong, Z., Chen, D.: Retrieval-based language models and applications. In: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 6: Tutorial Abstracts). pp. 41–46 (2023) 2

[3] Asai, A., Schick, T., Lewis, P., Chen, X., Izacard, G., Riedel, S., Hajishirzi, H., Yih, W.t.: Task-aware retrieval with instructions. Findings of ACL (2022) 7, 14

[4] Blattmann, A., Rombach, R., Oktay, K., Müller, J., Ommer, B.: Semi-parametric neural image synthesis. In: Oh, A.H., Agarwal, A., Belgrave, D., Cho, K. (eds.) Advances in Neural Information Processing Systems (2022) 2

[5] Brooks, T., Holynski, A., Efros, A.A.: Instructpix2pix: Learning to follow image editing instructions. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 18392–18402 (2023) 14

[6] Brooks, T., Peebles, B., Holmes, C., DePue, W., Guo, Y., Jing, L., Schnurr, D., Taylor, J., Luhman, T., Luhman, E., Ng, C., Wang, R., Ramesh, A.: Video generation models as world simulators (2024), https://openai.com/research/video-generation-models-as-world-simulators 14

[7] Chang, Y., Narang, M., Suzuki, H., Cao, G., Gao, J., Bisk, Y.: Webqa: Multihop and multimodal qa. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 16495–16504 (2022) 3, 7

[8] Changpinyo, S., Pont-Tuset, J., Ferrari, V., Soricut, R.: Telling the what while pointing to the where: Multimodal queries for image retrieval. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 12136–12146 (2021) 14

[9] Chen, W., Hu, H., Saharia, C., Cohen, W.W.: Re-imagen: Retrieval-augmented text-to-image generator. The International Conference on Learning Representations (2022) 2

[10] Chen, Y., Hu, H., Luan, Y., Sun, H., Changpinyo, S., Ritter, A., Chang, M.W.: Can pre-trained vision and language models answer visual information-seeking questions? Proceedings of Conference on Empirical Methods in Natural Language Processing (2023) 3, 7

[11] Chen, Y.C., Li, L., Yu, L., El Kholy, A., Ahmed, F., Gan, Z., Cheng, Y., Liu, J.: Uniter: Learning universal image-text representations (2019) 8

[12] Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H.W., Sutton, C., Gehrmann, S., et al.: Palm: Scaling language modeling with pathways. arXiv preprint arXiv:2204.02311 (2022) 14

[13] Chung, H.W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., Li, Y., Wang, X., Dehghani, M., Brahma, S., et al.: Scaling instruction-finetuned language models. arXiv preprint arXiv:2210.11416 (2022) 7, 14

[14] Dai, W., Li, J., Li, D., Tiong, A.M.H., Zhao, J., Wang, W., Li, B.A., Fung, P., Hoi, S.C.H.: Instructblip: Towards general-purpose vision-language models with instruction tuning. Advances in neural information processing systems (2023) 14

[15] Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. The International Conference on Learning Representations (2020) 9

[16] Fu*, S., Tamir*, N., Sundaram*, S., Chai, L., Zhang, R., Dekel, T., Isola, P.: Dreamsim: Learning new dimensions of human visual similarity using synthetic data. Advances in neural information processing systems (2023) 3, 7

[17] Ge, Y., Zhao, S., Zeng, Z., Ge, Y., Li, C., Wang, X., Shan, Y.: Making llama see and draw with seed tokenizer. arXiv preprint arXiv:2310.01218 (2023) 14

[18] Girdhar, R., El-Nouby, A., Liu, Z., Singh, M., Alwala, K.V., Joulin, A., Misra, I.: Imagebind: One embedding space to bind them all. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 15180–15190 (2023) 14

[19] Han, X., Wu, Z., Huang, P.X., Zhang, X., Zhu, M., Li, Y., Zhao, Y., Davis, L.S.: Automatic spatially-aware fashion concept discovery. In: Proceedings of the IEEE International Conference on Computer Vision (2017) 3, 7

[20] Hu, H., Chan, K.C., Su, Y.C., Chen, W., Li, Y., Sohn, K., Zhao, Y., Ben, X., Gong, B., Cohen, W., et al.: Instruct-imagen: Image generation with multi-modal instruction. arXiv preprint arXiv:2401.01952 (2024) 14

[21] Hu, H., Luan, Y., Chen, Y., Khandelwal, U., Joshi, M., Lee, K., Toutanova, K., Chang, M.W.: Open-domain visual entity recognition: Towards recognizing millions of wikipedia entities. Proceedings of the IEEE International Conference on Computer Vision (2023) 3, 4, 6

[22] Jain, A., Guo, M., Srinivasan, K., Chen, T., Kudugunta, S., Jia, C., Yang, Y., Baldridge, J.: Mural: multimodal, multitask retrieval across languages. Findings of the Association for Computational Linguistics: EMNLP (2021) 13

[23] Jia, C., Yang, Y., Xia, Y., Chen, Y.T., Parekh, Z., Pham, H., Le, Q., Sung, Y.H., Li, Z., Duerig, T.: Scaling up visual and vision-language representation learning with noisy text supervision. In: International Conference on Machine Learning. pp. 4904–4916. PMLR (2021) 13

[24] Johnson, J., Douze, M., Jégou, H.: Billion-scale similarity search with GPUs. IEEE Transactions on Big Data **7**(3), 535–547 (2019) 8

[25] Karpathy, A., Fei-Fei, L.: Deep visual-semantic alignments for generating image descriptions. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 3128–3137 (2015) 13

[26] Kim, W., Son, B., Kim, I.: Vilt: Vision-and-language transformer without convolution or region supervision. In: International Conference on Machine Learning. pp. 5583–5594. PMLR (2021) 13
[27] Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. International Conference on Machine Learning (2023) 8
[28] Li, J., Li, D., Xiong, C., Hoi, S.: Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In: International Conference on Machine Learning. pp. 12888–12900. PMLR (2022) 2, 4, 7, 8
[29] Li, J., Selvaraju, R., Gotmare, A., Joty, S., Xiong, C., Hoi, S.C.H.: Align before fuse: Vision and language representation learning with momentum distillation. Advances in neural information processing systems **34**, 9694–9705 (2021) 13
[30] Lin, T.Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., Zitnick, C.L.: Microsoft coco: Common objects in context. In: Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13. pp. 740–755. Springer (2014) 2, 3, 7
[31] Lin, X.V., Chen, X., Chen, M., Shi, W., Lomeli, M., James, R., Rodriguez, P., Kahn, J., Szilvasy, G., Lewis, M., et al.: Ra-dit: Retrieval-augmented dual instruction tuning. arXiv preprint arXiv:2310.01352 (2023) 14
[32] Liu, F., Wang, Y., Wang, T., Ordonez, V.: Visual news: Benchmark and challenges in news image captioning. Proceedings of the Conference on Empirical Methods in Natural Language Processing (2021) 3, 7
[33] Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. Advances in neural information processing systems (2023) 14
[34] Liu, S., Feng, W., Chen, W., Wang, W.Y.: Edis: Entity-driven image search over multimodal web content. Proceedings of Conference on Empirical Methods in Natural Language Processing (2023) 2, 3, 4, 6
[35] Liu, Z., Xiong, C., Lv, Y., Liu, Z., Yu, G.: Universal vision-language dense retrieval: Learning a unified representation space for multi-modal retrieval. In: The Eleventh International Conference on Learning Representations (2022) 4
[36] Liu, Z., Rodriguez-Opazo, C., Teney, D., Gould, S.: Image retrieval on real-life images with pre-trained vision-and-language models. In: Proceedings of the IEEE/CVF International Conference on Computer Vision. pp. 2125–2134 (2021) 3, 6
[37] Luo, M., Fang, Z., Gokhale, T., Yang, Y., Baral, C.: End-to-end knowledge retrieval with multi-modal queries. Annual Meeting of the Association for Computational Linguistics (2023) 14
[38] Mishra, S., Khashabi, D., Baral, C., Hajishirzi, H.: Cross-task generalization via natural language crowdsourcing instructions. In: Annual Meeting of the Association for Computational Linguistics (2021) 14
[39] OpenAI: Gpt-4 technical report. ArXiv **abs/2303.08774** (2023), https://api.semanticscholar.org/CorpusID:257532815 14

[40] Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al.: Training language models to follow instructions with human feedback. Advances in Neural Information Processing Systems **35**, 27730–27744 (2022) 14
[41] Pan, X., Dong, L., Huang, S., Peng, Z., Chen, W., Wei, F.: Kosmos-g: Generating images in context with multimodal large language models. arXiv preprint arXiv:2310.02992 (2023) 14
[42] Plummer, B.A., Wang, L., Cervantes, C.M., Caicedo, J.C., Hockenmaier, J., Lazebnik, S.: Flickr30k entities: Collecting region-to-phrase correspondences for richer image-to-sentence models. In: Proceedings of the IEEE international conference on computer vision. pp. 2641–2649 (2015) 13
[43] Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PMLR (2021) 2, 4, 7, 8
[44] Ram, O., Levine, Y., Dalmedigos, I., Muhlgay, D., Shashua, A., Leyton-Brown, K., Shoham, Y.: In-context retrieval-augmented language models. Transactions of the Association for Computational Linguistics (2023) 2
[45] Sheynin, S., Ashual, O., Polyak, A., Singer, U., Gafni, O., Nachmani, E., Taigman, Y.: kNN-diffusion: Image generation via large-scale retrieval. In: The Eleventh International Conference on Learning Representations (2023) 2
[46] Singhal, A., et al.: Modern information retrieval: A brief overview. IEEE Data Eng. Bull. **24**(4), 35–43 (2001) 2
[47] Su, H., Shi, W., Kasai, J., Wang, Y., Hu, Y., Ostendorf, M., Yih, W.t., Smith, N.A., Zettlemoyer, L., Yu, T.: One embedder, any task: Instruction-finetuned text embeddings. In: Rogers, A., Boyd-Graber, J., Okazaki, N. (eds.) Findings of the Association for Computational Linguistics: ACL 2023 (Jul 2023) 14
[48] Sun, Q., Cui, Y., Zhang, X., Zhang, F., Yu, Q., Luo, Z., Wang, Y., Rao, Y., Liu, J., Huang, T., et al.: Generative multimodal models are in-context learners. arXiv preprint arXiv:2312.13286 (2023) 14
[49] Tang, Z., Yang, Z., Khademi, M., Liu, Y., Zhu, C., Bansal, M.: Codi-2: In-context, interleaved, and interactive any-to-any generation. arXiv preprint arXiv:2311.18775 (2023) 14
[50] Tang, Z., Yang, Z., Zhu, C., Zeng, M., Bansal, M.: Any-to-any generation via composable diffusion. Advances in Neural Information Processing Systems **36** (2024) 8, 14
[51] Team, G., Anil, R., Borgeaud, S., Wu, Y., Alayrac, J.B., Yu, J., Soricut, R., Schalkwyk, J., Dai, A.M., Hauth, A., et al.: Gemini: a family of highly capable multimodal models. arXiv preprint arXiv:2312.11805 (2023) 14
[52] Thakur, N., Reimers, N., Rücklé, A., Srivastava, A., Gurevych, I.: Beir: A heterogenous benchmark for zero-shot evaluation of information retrieval models. Advances in neural information processing systems (2021) 14
[53] Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al.: Llama: Open

and efficient foundation language models. arXiv preprint arXiv:2302.13971 (2023) 14
[54] Wang, B., Ping, W., McAfee, L., Xu, P., Li, B., Shoeybi, M., Catanzaro, B.: Instructretro: Instruction tuning post retrieval-augmented pretraining. arXiv preprint arXiv:2310.07713 (2023) 14
[55] Wei, J., Bosma, M., Zhao, V.Y., Guu, K., Yu, A.W., Lester, B., Du, N., Dai, A.M., Le, Q.V.: Finetuned language models are zero-shot learners. The International Conference on Learning Representations (2021) 7, 14
[56] Wu, H., Gao, Y., Guo, X., Al-Halah, Z., Rennie, S., Grauman, K., Feris, R.: 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. pp. 11307–11317 (2021) 3, 6, 7
[57] Wu, S., Fei, H., Qu, L., Ji, W., Chua, T.S.: Next-gpt: Any-to-any multimodal llm. arXiv preprint arXiv:2309.05519 (2023) 8, 14
[58] Xu, Z., Shen, Y., Huang, L.: Multiinstruct: Improving multi-modal zero-shot learning via instruction tuning. Annual Meeting of the Association for Computational Linguistics (2023) 14
[59] Yang, Z., Li, L., Lin, K., Wang, J., Lin, C.C., Liu, Z., Wang, L.: The dawn of lmms: Preliminary explorations with gpt-4v (ision). arXiv preprint arXiv:2309.17421 **9**(1), 1 (2023) 8
[60] Yasunaga, M., Aghajanyan, A., Shi, W., James, R., Leskovec, J., Liang, P., Lewis, M., Zettlemoyer, L., Yih, W.t.: Retrieval-augmented multimodal language modeling (2023) 2
[61] Yu, L., Shi, B., Pasunuru, R., Muller, B., Golovneva, O., Wang, T., Babu, A., Tang, B., Karrer, B., Sheynin, S., et al.: Scaling autoregressive multi-modal models: Pretraining and instruction tuning. arXiv preprint arXiv:2309.02591 (2023) 14
[62] Zhai, X., Mustafa, B., Kolesnikov, A., Beyer, L.: Sigmoid loss for language image pre-training. Proceedings of the IEEE International Conference on Computer Vision (2023) 8
[63] Zhang, K., Mo, L., Chen, W., Sun, H., Su, Y.: Magicbrush: A manually annotated dataset for instruction-guided image editing. Advances in neural information processing systems (2023) 14