The University of Manchester Research

SciMMIR

Link to publication record in Manchester Research Explorer

Citation for published version (APA):
Wu, S., Li, Y., Zhu, K., Zhang, G., Liang, Y., Ma, K., Xiao, C., Zhang, H., Yang, B., Chen, W., Huang, W., Moubayed, N. A., Fu, J., & Lin, C. (2024). *SciMMIR: Benchmarking Scientific Multi-modal Information Retrieval*.

Citing this paper
Please note that where the full-text provided on Manchester Research Explorer is the Author Accepted Manuscript or Proof version this may differ from the final Published version. If citing, it is advised that you check and use the publisher's definitive version.

General rights
Copyright and moral rights for the publications made accessible in the Research Explorer are retained by the authors and/or other copyright owners and it is a condition of accessing publications that users recognise and abide by the legal requirements associated with these rights.

Takedown policy
If you believe that this document breaches copyright please refer to the University of Manchester's Takedown Procedures [http://man.ac.uk/04Y6Bo] or contact uml.scholarlycommunications@manchester.ac.uk providing relevant details, so we can investigate your claim.

Download date:24. May. 2024

SciMMIR: Benchmarking Scientific Multi-modal Information Retrieval

Siwei Wu[,1] **Yizhi Li**[,1] **Kang Zhu** **Ge Zhang**[,2] **Yiming Liang**[3]
Kaijing Ma[4] **Chenghao Xiao**[,5] **Haoran Zhang**[,4] **Bohao Yang**[1]
Wenhu Chen[,2] **Wenhao Huang**[,3] **Noura Al Moubayed**[5] **Jie Fu**[,4*] **Chenghua Lin**[,1*]
Multimodal Art Projection Research Community [1]University of Manchester [2]University of Waterloo
[3]01.ai [4]Hong Kong University of Science and Technology [5]Durham University

arXiv:2401.13478v1 [cs.IR] 24 Jan 2024

Abstract

Multi-modal information retrieval (MMIR) is a rapidly evolving field, where significant progress, particularly in image-text pairing, has been made through advanced representation learning and cross-modality alignment research. However, current benchmarks for evaluating MMIR performance in image-text pairing within the scientific domain show a notable gap, where chart and table images described in scholarly language usually do not play a significant role. To bridge this gap, we develop a specialised scientific MMIR (SciMMIR) benchmark by leveraging open-access paper collections to extract data relevant to the scientific domain. This benchmark comprises 530K meticulously curated image-text pairs, extracted from figures and tables with detailed captions in scientific documents. We further annotate the image-text pairs with two-level subset-subcategory hierarchy annotations to facilitate a more comprehensive evaluation of the baselines. We conducted zero-shot and fine-tuning evaluations on prominent multi-modal image-captioning and visual language models, such as CLIP and BLIP. Our analysis offers critical insights for MMIR in the scientific domain, including the impact of pre-training and fine-tuning settings and the influence of the visual and textual encoders. All our data and checkpoints are publicly available [1].

1 Introduction

Information retrieval (IR) systems are expected to provide a matched piece of information from an enormous but organised data collection according to given user queries. With the advancement of representation learning (Bengio et al., 2013), the methodological paradigm of IR systems has evolved from using lexical matching to retrieve textual data (Luhn, 1957; Jones et al., 2000; Robertson et al., 2009) to a mixture fashion of similarity matching in a learned representation space, which supports additional modalities such as images and audios other than texts (Karpukhin et al., 2020; Chen et al., 2020b; Koepke et al., 2022). Whilst enabling broader application scenarios, such multi-modal information retrieval (MMIR) systems also introduce new challenges in evaluation. Although previous studies have evaluated image-text retrieval across general topics using large-scale paired datasets from sources such as Wikipedia (Young et al., 2014; Lin et al., 2014; Srinivasan et al., 2021; Luo et al., 2023), there is a notable gap in comprehensively assessing MMIR models within the scientific domain. In this domain, unique challenges arise from the complex and dense semantics of scientific images and the sophisticated language preferences of researchers. For example, current MMIR models often ignore aspects such as learning from histograms or plot figures and lack pre-training data necessary for effectively extracting key textual information from table images.

To fill such a gap, we introduce **SciMMIR**, a **S**cientific **M**ulti-**M**odal **I**nformation **R**etrieval benchmark to evaluate models' MMIR ability in the scientific domain. SciMMIR is built upon a dataset of 530K scientific image-text pairs with sub-class annotations. We collect the figures, tables, and their associated captions from scholarly documents on arXiv[2], an open-access archival collection, to construct image-text pairs. To comprehensively evaluate the cross-modality aligned representations learned by the models, our SciMMIR benchmark defines the retrieval tasks in *two direction retrieval*, including searching the matched textual caption within the candidate pool with a given image (`img`→`txt`), and finding the corresponding figure or table image with a caption (`txt`→`img`).

*Corresponding authors.
[1]https://github.com/Wusiwei0410/SciMMIR
[2]https://arxiv.org

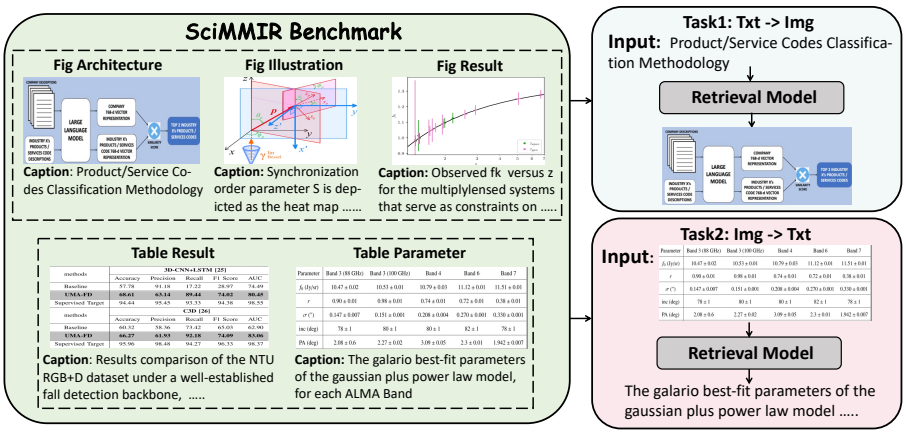

Figure 1: The Illustration of SciMMIR Framework.

Furthermore, we conduct *finely-grained subset evaluation* because we contend that analysing baseline performances can be enhanced by evaluating them on subsets characterised by specific attributes. For example, by limiting evaluations to searching for matches only within "figures describing architectures", models demonstrating poorer performance can be identified and improved specifically in their chart and shape modelling capabilities, rather than focusing on the recognition of letters and digits. Therefore, we annotate and categorise the image-text pairs into three figure-caption and two table-caption subcategories based on their distinctive characteristics. Models are required to retrieve information in a more detailed, finer-grained setting that could more comprehensively expose the shortcomings, as opposed to retrieving from all available candidates.

In our evaluation, we conducted meticulous experiments in both zero-shot and fine-tuning settings across various subcategories. These were performed on the chosen image captioning models and visual language models (VLMs) to enrich the corpus of knowledge in future MMIR research. We list the key takeaway insights as follows:

1. Our findings reveal that the pre-training tasks and the dataset selection play significant roles in determining the performance in both scenarios of SciMMIR. An expected observation is that after fine-tuning with data specific to the scientific domain, there would be a marked performance improvement for both `txt`→`img` and `img`→`txt` tasks, underlining the effectiveness of domain-specific adaptation.
2. The results suggest a nuanced distinction between the MMIR tasks involving the figure and table subsets as well. The performance on the figure subset could be effectively improved by a scientific data domain adaption, showing the generalisability of the visual encoders. In contrast, metrics on the table subset remain less promising, due to that table images seldom appear in the scope of the image-text pre-training dataset, and such kind of task is more sensitive to the captured visually presented textual information.
3. Regardless of parameter size, the BLIP2 series models generally perform better among the pre-trained VLMs. The extra zero-shot multi-modal information retrieval capability may be brought by the distinct pre-training tasks including image-text matching and image-text contrastive learning, other than language modelling.

These findings underscore the importance of tailored approaches for different data types within the scientific MMIR framework. A more in-depth exploration of these findings are given in §5.

To conclude, our contributions are:

- providing the first **benchmark** for scientific multi-modal information retrieval models;
- releasing a public 530K scientific image-text **dataset** with fine-grained annotations;
- comprehensively **analysing** performances of prevalent multi-modal information retrieval models.

2 Related Work

General Information Retrieval Information Retrieval has lied at the core of NLP, and has been facilitated by dense representation learning in the

past few years (Reimers and Gurevych, 2019; Karpukhin et al., 2020). More recently, unified representations across tasks have become a consensus of some, and this line of research proposes to understand and evaluate task-agnostic representation in a single representation space (Muennighoff et al., 2023; Asai et al., 2022; Su et al., 2022; Wei et al., 2023). In another vein, domain generalisation has always been seen as a key weakness to address for IR models (Thakur et al., 2021). Through the subpar performance of general domain methods on SciMMIR, we will present that scientific IR, especially multi-modal one, remains an OOD task and domain, despite the advancement of techniques in general information retrieval.

Multi-modal Information Retrieval In earlier multi-modal representation learning research, small-scale cross-modal retrieval datasets including MSCOCO (Lin et al., 2014) and Flickr30k (Plummer et al., 2015) have facilitated the alignment between visual and language representation. Efforts have since shifted towards large-scale vision-language pretraining (Radford et al., 2021; Kim et al., 2021; Li et al., 2021; Jia et al., 2021; Yu et al., 2022), and these small-scale retrieval datasets, in turn, become the standard evaluation of such systems. The advancements in multi-modal representation alignment have also facilitated multi-modal retrieval-augmented generation (Chen et al., 2022; Yasunaga et al., 2022; Hu et al., 2023; Lin et al., 2023). More recently, evaluating unified cross-modal representation across diverse tasks has emerged as a prevalent trend (Wei et al., 2023).

Scientific Document Learning Scientific information retrieval has been moderately evaluated in NLP, with SciFact (Wadden et al., 2020) and SCIDOCS (Cohan et al., 2020) commonly incorporated in popular zero-shot information retrieval benchmarks (Thakur et al., 2021). More complex tasks are proposed in this area, such as DORIS-MAE, a task to retrieve documents in response to complex, multifaceted scientific queries (Wang et al., 2023). In the multi-modal area, VQA (Antol et al., 2015) has been another major approach to evaluating vision-language systems, concerning in-depth visual grounding, instead of distributional priors (Agrawal et al., 2018). This is where work with a similar scope to ours in the scientific domain such as PlotQA and ChartQA is seen (Methani et al., 2020; Masry et al., 2022). SciMMIR distinguishes by offering extensive coverage across diverse figure and table subcategories, a larger dataset size, and the utilisation of real-world data that is naturally paired and not reliant on human annotation.

Subset	Subcategory	Number		
		Train	Valid	Test
Figure	Result	296,191	9,676	9,488
	Illustration	46,098	1,504	1,536
	Architecture	13,135	447	467
Table	Result	126,999	4,254	4,229
	Parameter	15,856	552	543
Total		498,279	16,433	16,263

Table 1: Statistics of the SciMMIR dataset.

3 Dataset Collection

Dataset Overview As shown in Table 1, the SciMMIR dataset comprises 530K samples, and the average length of captions in the dataset is 43.22 words. The dataset is split as train, valid, and test sets with $498,279$, $16,433$, and $16,263$ samples, respectively.

Data Annotation After in-depth observation and statistics of *Figure* and *Table* data presented in various scientific papers, we define a data hierarchical architecture of "Two subsets, Five subcategories" in the SciMMIR benchmark. Based on the Figure and Table subsets, a finer-grained division is carried out, as shown in Table 2. We use a set of specific extracted keywords to classify the dataset using the title and caption of each sample. The subset and subcategory classification results are shown in Table 1, providing a structured and standardised basis for subsequent experiments.

Subset	Subcategory	Description
Figure	Architecture	Depicts scientific study frameworks and conceptual designs.
	Illustration	Illustrates complex scientific concepts or data relationships.
	Result	Visually presents scientific research outcomes.
Table	Parameter	Details of key parameters and variables in studies.
	Result	Summarises and displays experiment/study results.

Table 2: The hierarchical architecture for SciMMIR.

4 Experiment

4.1 Retrieval Baseline

We evaluate a wide range of baseline models. Drawing on the distributional gap between the scientific and general domains highlighted previously, we further illustrate the relationship between multi-modal information retrieval performance and the distributions already learned by the models. To this end, we collect the information of the pre-training phase for the baseline models in Table 3 and references in Appendix A.

Image Captioning Models We utilise image captioning models as a set of baselines, including **CLIP-base** (Radford et al., 2021) and **BLIP-base** (Li et al., 2022), that have particularly learned the pairing relationship between corresponding image and text with strong supervision signal. We evaluate these image captioning models trained with general domain datasets in both zero-shot and fine-tuning settings to investigate the need for scientific domain adaption. We also introduce **BERT** (Devlin et al., 2018) as an alternative text encoder for the captioning models (denoted as "+BERT" in the tables), where such ensemble baselines might reveal the influence of the text encoders.

Visual Language Models Additionally, we select large visual language models (VLMs) trained for multi-modal tasks like visual question answering to examine their zero-shot MMIR performances.

- **BLIP2** (Li et al., 2023) series model uses a querying transformer module to address the modality gap. We chose the models grounded in large language models (LLMs), BLIP2-OPT-2.7B, BLIP2-OPT-6.7B, BLIP2-FLAN-T5-XL and BLIP2-FLAN-T5-XXL, as our baselines.
- **Fuyu-8B**[3] is a multi-modal decoder-only transformer for both image and text modelling that directly projects image patches into the text embedding space.
- **LLaMA-Adapter2-7B** (Gao et al., 2023) efficiently fine-tunes additional parameters based on the LLaMA model (Touvron et al., 2023), where the extra expert models further boost its image understanding capability.
- **Kosmos-2** (Peng et al., 2023) aligns perception with language and adds the ability to recognise and understand images based on its multi-turn dialogue and reasoning capabilities. Specifically, it achieves the capability of grounding images, allowing it to interact with inputs at the object level.
- **mPLUGw-OWL2** (Ye et al., 2023) introduces a Modality-Adaptive Module (MAM) module into the large language model. By adding a small number of parameters during the attention process, it further learns a shared space for both vision and language representations.

[3] https://www.adept.ai/blog/fuyu-8b

4.2 Evaluation Protocol

Task Definition The SciMMIR benchmark designs two directions of multi-modal retrieval tasks. To be specific, it has a forward direction retrieval task and an inverse direction retrieval task:

- **`txt`→`img`**: the forward direction retrieval task, given the text which is relevant to an image, retrieves the image in the candidate set.
- **`img`→`txt`**: the inverse direction retrieval task, given the image which is relevant to a text, retrieves the text in the candidate set.

The relevance score in retrieval ranking is defined as the dot product between the visual and textual representations. In addition to assessing the model's performance on the overall test set (defined as "ALL" in tables), we evaluate the retrieval models in different subsets and subcategories to scrutinize their abilities. In more detail, we assess the model's performance on various fine-grained subcategories of the test set, including Figure Architecture, Figure Illustration, Figure Result, Table Result and Table Parameter, as well as the performance on the Figure and Table subset.

Metrics In this paper, we use MRR and Hits@K metrics to assess the information retrieval models' ability in SciMMIR benchmark.

- **MRR** stands for Mean Reciprocal Rank, and it is calculated by the reciprocal of the golden label's ranking in candidates. A higher MRR score indicates better performance.
- **Hits@K** assesses the accuracy of the retrieval system by checking whether the golden label is present within the top-k ranked results. Hits@10 are used in our measurement.

Model	Pre-training Data		Pre-training Task	Trainable & *Frozen Parameters		
	Domain	Number		Visual	Textual	Align
CLIP-base	Internet Crawled	400M	Contrastive	62M	63M	/
BLIP-base	COCO, VG, CC3M, CC12M, SBU, LAION-400M	129M	Image-Text Contrastive, Image-Text Matching, Language Modeling	25.5M	108M	/
BLIP2-OPT-2.7B	COCO, VG, CC3M, CC12M, SBU, LAION-400M	129M	Image-Text Contrastive, Image-Text Matching, Image-grounded Text Generation	*1.3B	*2.7B	*2.7B
BLIP2-OPT-6.7B					*6.7B	*6.7B
BLIP2-FLAN-T5-XL					*2.85B	*2.85B
BLIP2-FLAN-T5-XXL					*11.3B	*11.3B
LLaMA-Adapter2-7B	LAION-400M, COYO, MMC4, SBU, CC3M, COCO	56.7M	Fine-Tuning only	*62M	*7B	14M
Kosmos-2	GRIT	90M	Language Modeling	0.3B	1.3B	19M
mPLUGw-OWL2	COCO, CC3M, CC12M, LAION-5B, COYO, DataComp	400M	Language Modeling	0.3B	7B	0.9B
Fuyu-8B	/	/	language modelling	8.3B		/

Table 3: The pre-training information of the baselines. "_" refers to non-public or not fully public data.

	Model	ALL				Figure*				Table*			
		txt→img		img→txt		txt→img		img→txt		txt→img		img→txt	
		MRR	Hits@10	MRR	Hits@10	MRR	Hits@10	MRR	Hits@10	MRR	Hits@10	MRR	Hits@10
FT	CLIP-base	23.77	38.76	23.76	38.63	26.54	42.80	26.56	42.96	17.11	29.05	17.02	28.19
	CLIP-base+BERT	14.86	28.38	16.02	30.20	16.91	31.81	18.58	34.26	9.94	20.17	9.86	20.45
	BLIP-base	24.22	42.14	24.26	41.82	26.35	45.08	26.33	44.64	19.18	35.20	19.34	35.17
	BLIP-base+BERT	**34.53**	**55.41**	**35.43**	**55.78**	**37.48**	**59.07**	**38.26**	**59.35**	**27.55**	**46.80**	**28.77**	**47.37**
ZS	CLIP-base	2.30	3.79	1.85	3.40	2.54	4.25	2.03	3.67	1.68	2.64	1.37	2.70
	BLIP-base	0.07	0.09	0.08	0.07	0.09	0.12	0.06	0.04	0.02	0.02	0.15	0.12
	BLIP2-FLAN-T5-XL	0.46	0.82	0.16	0.21	0.55	1.02	0.14	0.18	0.22	0.29	0.21	0.27
	BLIP2-FLAN-T5-XXL	0.68	1.29	0.11	0.13	0.89	1.71	0.13	0.15	0.17	0.23	0.09	0.09
	BLIP2-OPT-2.7B	0.61	1.02	0.25	0.40	0.60	0.97	0.31	0.48	0.61	1.16	0.12	0.21
	BLIP2-OPT-6.7B	0.05	0.06	0.06	0.05	0.06	0.06	0.02	0.00	0.05	0.06	0.14	0.17
	Fuyu-8B	0.06	0.03	0.07	0.06	0.05	0.02	0.03	0.02	0.07	0.06	0.15	0.15
	mPLUG-Owl2-LLaMA2-7B	0.26	0.42	0.09	0.07	0.34	0.56	0.06	0.03	0.05	0.06	0.15	0.17
	Kosmos-2	0.11	0.15	0.09	0.14	0.15	0.22	0.10	0.17	0.01	0.00	0.06	0.06
	LLaMA-Adapter2-7B	0.06	0.06	0.05	0.07	0.06	0.08	0.06	0.10	0.04	0.02	0.03	0.00

Table 4: The main results of SciMMIR benchmark. * refers to average metrics grouped by the subcategories in the Figure and Table subsets.

Zero-shot We provide a zero-shot (ZS) setting in the evaluation for all baselines. For the *image-captioning* models, the learned features extracted by the visual encoder and textual encoder are directly used, since they have been aligned to the same representation space. For the *visual language* models, the visual representation remains the same but the representations from the textual module are used depending on their architectures. For encoder-decoder textual models such as BLIP2-FLAN-T5s, we use the output features from the encoder as the text features. For decoder-only textual models like BLIP2-OPTs, we take mean pooling of outputs from the last decoder layer.

Fine-tuning Other than that, we provide fine-tuning (FT) evaluation for the relatively smaller CLIP-base and BLIP-base models trained with our data. During the fine-tuning, we employ the standard contrastive learning approach (Chen et al., 2020a) to minimise the distance between positive text-image pairs and decrease the scores between negative text-image pairs within a batch of samples. In addition to training the models on the entire training set, we also train them on different subsets of the training data to investigate the modelling abilities in a fine-grained manner. In all experiments, we fine-tune the models for 5 epochs on an A100 GPU, with a learning rate of 2e-5.

5 Result Analysis

5.1 Overall Evaluation

Following the designed evaluation protocol, we show the baseline performances in the universal set (ALL), and averagely group the subcategory metrics in Figure and Table subset in Table 4. In this subsection, we mainly discuss the results regarding the two-direction retrieval tasks and the subset performance.

Model	Fig Architecture				Fig Illustration				Fig Result			
	txt→img		img→txt		txt→img		img→txt		txt→img		img→txt	
	MRR	Hits@10	MRR	Hits@10	MRR	Hits@10	MRR	Hits@10	MRR	Hits@10	MRR	Hits@10
CLIP-base	4.31	6.85	4.19	7.07	4.23	7.16	2.66	4.62	2.20	3.68	1.84	3.37
BLIP-base	0.08	0.21	0.04	0.00	0.08	0.13	0.05	0.07	0.09	0.11	0.06	0.04
BLIP2-FLAN-T5-XL	0.64	2.36	0.20	0.00	1.16	1.76	0.16	0.20	0.45	0.85	0.13	0.19
BLIP2-FLAN-T5-XLL	1.72	3.21	0.24	0.21	1.39	2.67	0.20	0.39	0.77	1.50	0.11	0.11
BLIP2-OPT-2.7B	0.43	0.64	0.08	0.21	0.87	1.37	0.22	0.26	0.57	0.92	0.33	0.53
BLIP2-OPT-6.7B	0.06	0.00	0.04	0.00	0.09	0.13	0.02	0.00	0.05	0.05	0.02	0.00
Fuyu-8B	0.03	0.00	0.02	0.00	0.08	0.07	0.04	0.07	0.05	0.01	0.03	0.01
Kosmos-2	0.69	0.64	0.43	0.64	0.20	0.59	0.21	0.39	0.12	0.12	0.14	0.07
mPLUG-Owl2-LLaMA2-7B	0.36	0.64	0.11	0.00	0.93	1.37	0.09	0.07	0.25	0.43	0.05	0.03
LLaMA-Adapter2-7B	0.05	0.00	0.04	0.00	0.14	0.13	0.06	0.07	0.05	0.07	0.06	0.11

Table 5: The zero-shot results of multimodal models on Figure subsets of our SciMMIR benchmark.

Model	Table Result				Table Parameter			
	txt→img		img→txt		txt→img		img→txt	
	MRR	Hits@10	MRR	Hits@10	MRR	Hits@10	MRR	Hits@10
CLIP-base	1.60	2.53	1.25	2.48	2.35	3.50	2.29	4.42
BLIP-base	0.02	0.02	0.16	0.14	0.01	0.00	0.07	0.00
BLIP2-FLAN-T5-XL	0.21	0.28	0.17	0.24	0.28	0.37	0.57	0.55
BLIP2-FLAN-T5-XLL	0.16	0.21	0.07	0.05	0.25	0.37	0.29	0.37
BLIP2-OPT-2.7B	0.60	1.21	0.12	0.19	0.67	0.74	0.16	0.37
BLIP2-OPT-6.7B	0.05	0.07	0.13	0.17	0.07	0.00	0.22	0.18
Fuyu-8B	0.07	0.07	0.11	0.12	0.07	0.00	0.48	0.37
Kosmos-2	0.01	0.00	0.05	0.05	0.01	0.00	0.13	0.18
mPLUG-Owl2-LLaMA2-7B	0.05	0.07	0.15	0.17	0.06	0.00	0.16	0.18
LLaMA-Adapter2-7B	0.04	0.02	0.03	0.00	0.02	0.00	0.04	0.00

Table 6: The zero-shot results of multi-modal models on Table subsets of our SciMMIR benchmark datasets.

For both the forward (`txt`→`img`) and inverse (`img`→`txt`) tasks, we find that small models fine-tuned with our in-domain scientific image-text data generally have superior performances in all settings of SciMMIR benchmark. As this shows the necessity of domain adaption for improvement in the SciMMIR task, our designed tasks remain challenging for most of the models. In the forward retrieval task, many of the zero-shot large VLMs demonstrate insufficient performance, with the MRR and Hits@10 metrics not surpassing 0.5% in the ALL setting. It is worth mentioning that the CLIP-base model is well-trained since its zero-shot `txt`→`img` performance is better than all other VLMs of dominant parameter sizes.

The performance of the fine-tuned multi-modal models in information retrieval involving both figures and tables is notably promising. However, the results indicate a significantly higher performance solely on the Figure subset compared to the Table subset, suggesting the unexplored challenges of multi-modal information retrieval for table images. The lower scores on Table subset could be due to the scarcity of table-style images in the pre-training dataset and the lack of textual information perception ability of the visual encoders.

5.2 Analysis on Zero-shot Setting

To provide a more thorough analysis, we show the zero-shot performances of the baselines across different subcategories in Table 5 and Table 6, where only the images or texts from the same subcategory would be considered as retrieval candidates.

CLIP-base and BLIP-base The CLIP-base captioning baseline, which is specifically designed for image-text matching, shows certain generalisability in both forward and inverse retrieval across all subcategories within the Figure and Table subsets. In contrast, the BLIP-base model shows nearly no signs of training on the scientific domain multi-modal data.

Zero-shot `txt`→`img` The selected large pre-trained VLMs do not perform well on various subcategories in both the Figure and Table subsets.

In the Table subset, all models except CLIP-base have relatively lower performance. In the Figure subset, the BLIP2-FLAN-T5 series models show slightly better performance in Fig Subset. This could be attributed to the fact that the encoder in text encoder-decoder architecture can capture better textual features.

Zero-shot `img`→`txt` For the Figure subset, the performance of all VLMs in the reverse direction

is slightly worse than that in the forward direction. This indicates that VLMs' image-grounded text generation task can enhance the model's performance in multimodal retrieval for the forward direction, while the performance in the reverse direction is comparatively poorer. For the Table subset, the performance of all models is similarly poor in both directions, indicating that most models do not consider Table-style data too much during the pre-training process.

The effect of text-image matching task As shown in the 5 and 6, the BLIP2-series models outperform other large VLMs in both Figure's and Table's subcategories, specially for forward direction task. We believe that this is because BLIP2 takes into account the text-image matching task during the pre-training process. Most VLMs primarily focus on the image-grounded text generation task. However, the BLIP2 model addresses this limitation by incorporating a text-matching task and an image-grounded text generation task during its pre-training process to better align textual and visual information. Specifically, BLIP2 includes a specialized alignment module q-former as an information bottleneck for text-image alignment, eliminating the need for additional textual input to align image representations. This allows BLIP2 to acquire specific textual and visual representations for carrying out text-image matching. The experimental results demonstrate that other models solely relying on image-grounded text generation tasks may not yield effective representations for multi-modal retrieval. Therefore, dedicated pre-training models for multi-modal retrieval still require a primary focus on the text-image matching task.

5.3 Analysis on Fine-tuning Setting

Overall Analysis As shown in the Table 7, we fine-tune the models using data of different categories and evaluate the performance regarding all testing samples as candidates. The results indicate that training the model only with data from a specific subcategory leads to a significant performance gap compared to the model fine-tuned with all the data. There are two main factors contributing to this. Firstly, the dataset size of a specific subcategory is relatively small. Secondly, there are significant differences in data distribution among different subcategories. When training the model using only data from a particular subcategory, the model might become sensitive to that specific subcategory's data, but its overall performance on the samples from other subcategories will be poorer.

Besides, the BLIP-base+BERT model performs the best among all fine-tuning settings, while the performance of the CLIP model decreases when its text encoder is replaced. The vanilla BLIP-base model performs slightly better than CLIP-base and BLIP-base model almost has no performance on the zero-shot setting, which shows that the BLIP model has excellent domain knowledge learning capabilities.

The Impact of Subcategory Training Data As shown in Table 8 and Table 9, we report the result only regarding the specific subcategory testing sample for the sake of comprehensively investigating the impact of different subcategory training data.

For the BLIP model, the model's improvement on specific test subcategories generally aligns with the subcategories used for training. Besides, the model trained using a specific subcategory under a specific subset as training data can bring performance improvements to other subcategories of the corresponding subset in the test. This demonstrates the effectiveness of our annotation classification strategy in accurately clustering data points. On the other hand, it indicates the domain gaps among different subcategories and the correlation between different subcategories.

As for CLIP, the models trained on different subcategories consistently performing best in the Fig Architecture subcategory. We believe this is because the CLIP model has demonstrated a certain level of performance on the SciMMIR dataset and possesses a certain understanding of the data distribution within it.

The model trained on Fig Result data demonstrates good performance across the entire Figure subset. One reason could be that the Fig Result subset has the largest training proportion (54.02%) and text documents with relatively longer average length (**52.93 words** for Fig Result's average text length compared to the dataset's overall average text length of **43.23 words**) in the training dataset. This has highlighted the impact of training dataset size and its length coverage of text (Xiao et al., 2023a), on the performance and generalisability of retrieval models.

5.4 Text Encoder Generalisability

To investigate the impact of text encoders on multimodal retrieval tasks, we experimented by substi-

Model	Training Dataset	txt→img		img→txt	
		MRR	Hits@10	MRR	Hits@10
CLIP-base	ALL	23.77	38.76	23.76	38.63
	Fig Architecture	7.59	12.81	8.11	13.62
	Fig Illustration	13.28	21.83	14.02	23.00
	Fig Result	20.24	32.44	20.26	32.78
	Table Parameter	6.07	9.96	6.19	10.24
	Table Result	10.68	17.78	7.10	12.65
CLIP-base+BERT	ALL	14.86	28.38	16.02	30.20
BLIP-base	ALL	24.22	42.14	24.26	41.82
	Fig Architecture	0.43	0.73	0.45	0.70
	Fig Illustration	1.01	1.79	1.08	2.04
	Fig Result	15.72	28.42	16.09	28.89
	Table Parameter	0.16	0.26	0.21	0.32
	Table Result	2.73	5.39	2.51	5.02
BLIP-base+BERT	ALL	34.53	55.41	35.43	55.78

Table 7: The results of fine-tuning models which are trained on different subsets of training data and all training data. We report the averaged results of them on *All* testing subsets of our SciMMIR benchmark.

Model	Training Data	Fig Architecture				Fig Illustration				Fig Result			
		txt→img		img→txt		txt→img		img→txt		txt→img		img→txt	
		MRR	Hits@10	MRR	Hits@10	MRR	Hits@10	MRR	Hits@10	MRR	Hits@10	MRR	Hits@10
CLIP-base	All	28.02	41.97	27.18	43.47	26.28	42.77	26.67	43.75	26.51	42.84	26.51	42.81
	Fig Architecture	13.95	22.48	14.26	22.27	9.50	16.08	10.37	16.99	8.34	14.07	8.73	14.66
	Fig Illustration	21.07	32.12	20.66	31.05	18.09	29.30	19.08	30.27	15.81	25.93	16.25	26.64
	Fig Result	26.54	38.97	27.01	41.54	25.21	39.65	24.90	39.52	24.68	39.57	24.62	39.83
	Table Parameter	9.19	14.99	8.88	15.42	6.82	10.68	8.13	13.02	6.11	9.95	6.04	10.20
	Table Result	12.47	20.77	10.41	16.92	10.29	17.45	9.07	15.43	9.15	15.46	6.13	11.00
CLIP-base+BERT	All	15.24	28.27	17.29	30.62	17.08	31.64	18.45	34.11	16.96	31.99	18.66	34.44
BLIP-base	All	20.22	34.26	21.51	34.90	24.46	42.64	23.23	41.60	26.91	45.94	27.03	45.54
	Fig Architecture	1.03	2.36	1.46	3.00	0.58	0.98	0.52	0.78	0.50	0.85	0.48	0.78
	Fig Illustration	1.26	3.21	1.77	4.50	2.21	4.10	2.82	5.27	1.24	2.21	1.18	2.18
	Fig Result	16.69	27.19	16.23	28.05	18.20	34.77	18.57	34.51	20.94	37.43	21.30	37.88
	Table Parameter	0.04	0.00	0.04	0.00	0.10	0.13	0.10	0.13	0.11	0.18	0.15	0.16
	Table Result	0.53	0.86	0.26	0.21	0.44	0.78	0.76	1.30	0.75	1.38	0.83	1.56
BLIP-base+BERT	All	31.46	47.54	30.75	48.61	34.03	54.43	34.26	54.56	38.28	60.3	39.21	60.57

Table 8: The results of Fine-tuning models on Figure subsets of our SciMMIR benchmark.

Model	Training Data	Table Result				Table Parameter			
		txt→img		img→txt		txt→img		img→txt	
		MRR	Hits@10	MRR	Hits@10	MRR	Hits@10	MRR	Hits@10
CLIP-base	All	17.25	29.18	17.12	28.38	16.01	27.99	16.23	26.70
	Fig Architecture	4.84	8.35	5.40	9.29	4.90	8.10	6.78	12.15
	Fig Illustration	5.88	10.24	7.14	12.44	6.31	10.50	8.67	14.18
	Fig Result	9.26	15.39	9.50	15.84	8.67	14.73	8.98	14.92
	Table Parameter	5.39	9.22	5.49	8.87	5.84	9.39	6.47	9.21
	Table Result	13.69	22.42	8.14	14.80	13.39	20.63	7.42	13.26
CLIP-base+BERT	All	9.65	19.63	9.64	20.31	12.22	24.49	11.64	21.55
BLIP-base	All	19.23	35.16	19.43	35.21	18.75	35.54	18.66	34.81
	Fig Architecture	0.20	0.24	0.21	0.24	0.17	0.18	0.58	0.74
	Fig Illustration	0.11	0.02	0.24	0.38	0.25	0.37	0.45	1.1
	Fig Result	4.45	8.58	4.92	9.1	4.58	8.66	4.91	10.87
	Table Parameter	0.23	0.28	0.26	0.43	0.64	2.03	1.43	3.13
	Table Result	7.69	15.37	6.76	13.74	7.09	14.55	5.69	12.34
BLIP-base+BERT	All	27.52	46.89	29.22	47.77	27.82	46.04	25.18	44.20

Table 9: The results of Fine-tuning models on Table subsets of our SciMMIR benchmark.

tuting the text encoders in both BLIP-base and CLIP-base models with BERT-base. As shown in Table 7, replacing the text encoder of BLIP with BERT results in a significant improvement, while replacing the text encoder of CLIP led to a decline in performance. In the ALL setting, the MRR and Hits@10 metrics of CLIP-base decrease by 8.91% and 10.38% respectively in the `txt→img` task, and those metrics also decreased by 7.74% and 8.43% in `img→txt`. Conversely, combining BLIP-base with BERT resulted in a significant performance improvement. In the ALL setting, the MRR and Hits@10 metrics of the `txt→img` increased by 10.31% and 13.27% respectively, while the numbers of improvement are of 11.17% and 13.96% for the `img→txt` task.

Img Dim	Model	Training Dataset	txt→img		img→txt	
			MRR	Hits@10	MRR	Hits@10
224	BLIP-base	ALL	12.78	25.49	13.13	25.39
		Fig Architecture	0.24	0.31	0.26	0.38
		Fig Illustration	1.19	2.32	1.13	2.35
		Fig Result	9.03	17.48	9.30	17.87
		Table Result	1.79	3.55	1.58	3.18
		Table Parameter	0.17	0.33	0.22	0.35
224	BLIP-base+BERT	ALL	18.96	35.45	19.72	35.85
384	BLIP-base	ALL	24.22	42.14	24.26	41.82
		Fig Architecture	0.43	0.73	0.45	0.70
		Fig Illustration	1.01	1.79	1.08	2.04
		Fig Result	15.72	28.42	16.09	28.89
		Table Result	2.73	5.39	2.51	5.02
		Table Parameter	0.16	0.26	0.21	0.30
384	BLIP-base+BERT	ALL	34.53	55.41	35.43	55.78

Table 10: The averaged results of fine-tuning BLIP with different preprocessing image dimensions on *ALL* testing candidates of our SciMMIR benchmark.

The reasons for the performance changes being opposite after replacing the text encoder with BERT in both CLIP and BLIP could be as follows:

The CLIP Case CLIP itself exhibits strong performance. The images in the training data of CLIP are obtained through keyword searches in Wikipedia, which contains a significant amount of popular science text. In the zero-shot setting, the performance of the CLIP model is far better than that of the BLIP model on our SciMMIR benchmark. This has highlighted the nature of CLIP as a representation model. With the uniformity promise of contrastive learning (Wang and Isola, 2020), we conjecture that due to the textual and visual embeddings are well-aligned in an isotropic space in the pre-training phase of CLIP, replacing the text encoder with a highly anisotropic vanilla text encoder BERT hinders the stable alignment with the already learned vision encoder (Xiao et al., 2023b). We hypothesise that freezing the vision encoder in early fine-tuning might help guiding the replaced language model.

The BLIP Case On the one hand, in comparison to CLIP, BLIP utilizes BERT as its text encoder during the pre-training phase. Therefore, when fine-tuning BLIP to adapt to our SciMMIR benchmark data distribution, replacing BLIP's text encoder with BERT in terms of model structure is consistent. This can also minimize the impact on the model's performance. On the other hand, the zero-shot results indicate that the BERT in BLIP may not have adapted well to the domain of scientific research papers. Additionally, after experiencing the pre-training phase of BLIP, the fine-tuned BERT may not be able to effectively be adapted to new domain. By fine-tuning BLIP with the vanilla BERT, it can better establish the connection between images and text in the domian of scientific research papers.

5.5 Effects of Visual Encoder Resolution

In Table 4 for overall results, we compare the fine-tuned BLIP with default image preprocess dimension 384 and the fine-tuned CLIP with the default image preprocess dimension 224, where the results are relatively close. To make a fairer comparison, we decrease the image process dimension of BLIP-base model from 384 to 224, same as CLIP-base to conduct SciMMIR evaluation, as described in Table 10.

It can be seen that the granularity of image processing has a significant impact on model performance. When using a lower preprocessing dimension, the performance of BLIP is significantly decreased in both `txt`→`img` and `img`→`txt` tasks, using all training data settings. The performance of the CLIP model, which uses the same image processing dimension, is almost double that of BLIP.

Furthermore, although replacing the text encoder of BLIP with BERT during training on lower-dimensional (224) image preprocessed data improved the performance of the model, there was still a significant gap compared to CLIP. However, when the text encoder of BLIP was replaced with BERT during training on higher-dimensional image preprocessed data, the performance of the model was far superior to both CLIP and CLIP+BERT. This suggests that certain image-text shared interactive information is stored in the visual representations, and higher image quality can help the models

Model	Testing Data	Fig Architecture		Fig Illustration		Fig Results		Table Results		Table Parameters	
		txt→img	img→txt	txt→img	img→txt	txt→img	img→txt	txt→img	img→txt	txt→img	img→txt
FT-CLIP-base	Fig Architecture	15.91	17.56	15.82	15.40	66.57	65.42	1.24	1.26	0.45	0.36
	Fig Illustration	4.47	4.83	24.15	24.33	70.04	69.82	1.06	0.87	0.28	0.15
	Fig Results	3.28	3.47	11.38	11.35	83.28	83.33	1.73	1.57	0.34	0.28
	Table Results	0.12	0.15	0.22	0.44	3.33	3.77	87.77	86.9	8.56	8.74
	Table Parameters	0.39	0.53	0.41	0.88	4.53	5.91	67.64	65.86	27.03	26.81
	All	2.84	3.04	9.45	9.51	58.13	58.27	26.23	25.83	3.36	3.35
ZS-CLIP-base	Fig Architecture	5.85	5.89	31.61	13.96	56.53	72.53	4.95	5.30	1.07	0.66
	Fig Illustration	2.02	2.35	32.43	14.88	61.56	76.89	3.49	5.25	0.49	0.63
	Fig Results	1.73	1.55	26.38	10.19	63.10	79.85	7.54	7.87	1.25	0.55
	Table Results	0.13	0.39	1.22	5.74	13.95	34.90	68.87	51.53	15.83	7.44
	Table Parameters	0.24	0.61	2.21	8.62	17.22	38.64	59.96	42.01	20.37	10.13
	All	1.41	1.41	19.75	9.53	48.45	66.29	24.78	20.09	5.60	2.67

Table 11: The error analysis of CLIP model on our SciMMIR benchmark. The FT- stands for fine-tuned model and ZS- stands for zero shot.

better establish the connection between image and text representations.

5.6 Error Analysis

For better analysis of the performances, we calculate the ratio of samples that are retrieved from wrong subcategories in the top 10 answers predicted by the fine-tuned CLIP and vanilla CLIP.

As shown in Table 11, due to the larger volume of data in the categories labeled as Fig Results and Table Results (58.00% and 26.16%), the model tends to predict samples from these categories as answers. From the comparison between zero-shot and fine-tuning, it can be observed that fine-tuned model leads to a decrease in the proportion of incorrect predictions across almost all categories.

Under All setting, the fine-tuned model's predictions on different subcategories in the entire test set are consistent with the proportions of each subcategory in the training data (where the proportions of various subcategories in the training data are: Fig Architecture: 2.64%, Fig Illustration: 9.25%, Fig Results: 59.44%, Tab Results: 25.48%, Tab Parameter: 3.18%). This indicates that the proportions of various subcategories in the training set affect the model's final predictions. The higher the proportion of a subcategory in the training set, the higher the proportion of predictions for that subcategory during the testing phase.

Compared with zero-shot results, the fine-tuned model shows the largest improvement in prediction accuracy on the Fig Architecture and Fig Results testing data. However, the increase in prediction accuracy on the Table subset after fine-tuning is not obvious, indicating that retrieving information from Tables still poses significant challenges.

6 Conclusion

In summary, we introduce a novel benchmark and a corresponding dataset designed to address the gap in evaluating multi-modal information retrieval (MMIR) models in the scientific domain. Additionally, we have annotated the images into fine-grained subcategories based on characteristics of the figures and tables to facilitate a more comprehensive evaluation and analysis. Our zero-shot and fine-tuning evaluations, conducted on extensive baselines within various subsets and subcategories, offer valuable insights for future research.

References

Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. 2018. Don't just assume; look and answer: Overcoming priors for visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4971–4980.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

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

Yoshua Bengio, Aaron Courville, and Pascal Vincent. 2013. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828.

Minwoo Byeon, Beomhee Park, Haecheon Kim, Sungjun Lee, Woonhyuk Baek, and Saehoon Kim. 2022. Coyo-700m: Image-text pair dataset. https://github.com/kakaobrain/coyo-dataset.

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

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020a. A simple framework for contrastive learning of visual representations. In *International conference on machine learning*, pages 1597–1607. PMLR.

Wenhu Chen, Hexiang Hu, Xi Chen, Pat Verga, and William Cohen. 2022. Murag: Multimodal retrieval-augmented generator for open question answering over images and text. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5558–5570.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020b. Uniter: Universal image-text representation learning. In *European conference on computer vision*, pages 104–120. Springer.

Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel S Weld. 2020. Specter: Document-level representation learning using citation-informed transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2270–2282.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. 2023. Datacomp: In search of the next generation of multimodal datasets. *arXiv preprint arXiv:2304.14108*.

Peng Gao, Jiaming Han, Renrui Zhang, Ziyi Lin, Shijie Geng, Aojun Zhou, Wei Zhang, Pan Lu, Conghui He, Xiangyu Yue, Hongsheng Li, and Yu Qiao. 2023. Llama-adapter v2: Parameter-efficient visual instruction model. *arXiv preprint arXiv:2304.15010*.

Ziniu Hu, Ahmet Iscen, Chen Sun, Zirui Wang, Kai-Wei Chang, Yizhou Sun, Cordelia Schmid, David A Ross, and Alireza Fathi. 2023. 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.

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

K Sparck Jones, Steve Walker, and Stephen E. Robertson. 2000. A probabilistic model of information retrieval: development and comparative experiments: Part 2. *Information processing & management*, 36(6):809–840.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen Tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020*, pages 6769–6781. Association for Computational Linguistics (ACL).

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

A Sophia Koepke, Andreea-Maria Oncescu, Joao Henriques, Zeynep Akata, and Samuel Albanie. 2022. Audio retrieval with natural language queries: A benchmark study. *IEEE Transactions on Multimedia*.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalantidis, Li-Jia Li, David A Shamma, et al. 2017. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International journal of computer vision*, 123:32–73.

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

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

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

Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. 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.

Weizhe Lin, Jinghong Chen, Jingbiao Mei, Alexandru Coca, and Bill Byrne. 2023. Fine-grained late-interaction multi-modal retrieval for retrieval augmented visual question answering. *arXiv preprint arXiv:2309.17133*.

Hans Peter Luhn. 1957. A statistical approach to mechanized encoding and searching of literary information. *IBM Journal of research and development*, 1(4):309–317.

Man Luo, Zhiyuan Fang, Tejas Gokhale, Yezhou Yang, and Chitta Baral. 2023. End-to-end knowledge retrieval with multi-modal queries. *arXiv preprint arXiv:2306.00424*.

Ahmed Masry, Xuan Long Do, Jia Qing Tan, Shafiq Joty, and Enamul Hoque. 2022. Chartqa: A benchmark for question answering about charts with visual and logical reasoning. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2263–2279.

Nitesh Methani, Pritha Ganguly, Mitesh M Khapra, and Pratyush Kumar. 2020. Plotqa: Reasoning over scientific plots. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1527–1536.

Niklas Muennighoff, Nouamane Tazi, Loic Magne, and Nils Reimers. 2023. Mteb: Massive text embedding benchmark. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 2006–2029.

Vicente Ordonez, Girish Kulkarni, and Tamara Berg. 2011. Im2text: Describing images using 1 million captioned photographs. *Advances in neural information processing systems*, 24.

Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. 2023. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*.

Bryan A Plummer, Liwei Wang, Chris M Cervantes, Juan C Caicedo, Julia Hockenmaier, and Svetlana Lazebnik. 2015. 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.

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

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992.

Stephen Robertson, Hugo Zaragoza, et al. 2009. The probabilistic relevance framework: Bm25 and beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. 2022. Laion-5b: An open large-scale dataset for training next generation image-text models. *Advances in Neural Information Processing Systems*, 35:25278–25294.

Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. 2021. Laion-400m: Open dataset of clip-filtered 400 million image-text pairs. *arXiv preprint arXiv:2111.02114*.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565.

Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. 2021. Wit: Wikipedia-based image text dataset for multimodal multilingual machine learning. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2443–2449.

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

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. Beir: A heterogeneous benchmark for zero-shot evaluation

of information retrieval models. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

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

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7534–7550.

Jianyou Wang, Kaicheng Wang, Xiaoyue Wang, Prudhviraj Naidu, Leon Bergen, and Ramamohan Paturi. 2023. Scientific document retrieval using multi-level aspect-based queries. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Tongzhou Wang and Phillip Isola. 2020. Understanding contrastive representation learning through alignment and uniformity on the hypersphere. In *International Conference on Machine Learning*, pages 9929–9939. PMLR.

Cong Wei, Yang Chen, Haonan Chen, Hexiang Hu, Ge Zhang, Jie Fu, Alan Ritter, and Wenhu Chen. 2023. Uniir: Training and benchmarking universal multimodal information retrievers. *arXiv preprint arXiv:2311.17136*.

Chenghao Xiao, Yizhi Li, G Hudson, Chenghua Lin, and Noura Al Moubayed. 2023a. Length is a curse and a blessing for document-level semantics. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 1385–1396.

Chenghao Xiao, Yang Long, and Noura Al Moubayed. 2023b. On isotropy, contextualization and learning dynamics of contrastive-based sentence representation learning. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12266–12283.

Michihiro Yasunaga, Armen Aghajanyan, Weijia Shi, Rich James, Jure Leskovec, Percy Liang, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. 2022. Retrieval-augmented multimodal language modeling. *arXiv preprint arXiv:2211.12561*.

Qinghao Ye, Haiyang Xu, Jiabo Ye, Ming Yan, Anwen Hu, Haowei Liu, Qi Qian, Ji Zhang, Fei Huang, and Jingren Zhou. 2023. mplug-owl2: Revolutionizing multi-modal large language model with modality collaboration. *CoRR*, abs/2311.04257.

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

Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. 2022. Coca: Contrastive captioners are image-text foundation models. *arXiv preprint arXiv:2205.01917*.

Wanrong Zhu, Jack Hessel, Anas Awadalla, Samir Yitzhak Gadre, Jesse Dodge, Alex Fang, Youngjae Yu, Ludwig Schmidt, William Yang Wang, and Yejin Choi. 2023. Multimodal C4: An open, billion-scale corpus of images interleaved with text. *arXiv preprint arXiv:2304.06939*.

A The Baseline Pre-training Datasets

We provide a reference list for the pre-training image-text dataset mentioned in Table 3. COCO (Lin et al., 2014), consists of over 200,000 images across various categories including people, animals, everyday objects, and indoor scenes. VG (Krishna et al., 2017) dataset consists of over 100,000 images and covers a diverse range of visual concepts, including objects, scenes, relationships between objects, and other contextual information within images. CC3M (Sharma et al., 2018) contains over 3.3 million of images paired with descriptive captions, covering a wide range of topics and scenes, and providing a mix of everyday scenes, objects, and activities. CC12M (Changpinyo et al., 2021) contains 12.4 million image-text pairs, which is 3 times larger in scale compared to CC3M with a higher diversity degree containing more instances of out-of-domain (OOD) visual concepts. SBU (Ordonez et al., 2011) contains over 1 million images with visually relevant captions. The dataset is designed to be large enough for reasonable image-based matches to a query and the captions are filtered to ensure they are visually descriptive and likely to refer to visual content. LAION-400M (Schuhmann et al., 2021) is an open dataset that consists of 400 million image-text pairs, their CLIP embeddings, and KNN indices for efficient similarity search. It includes image URLs, corresponding metadata, CLIP image embeddings, and various KNN indices for quick search. LAION-5B (Schuhmann et al., 2022) is an open, large-scale dataset that consists of 5.85 billion image-text pairs, with 2.32 billion pairs in English. COYO (Byeon et al., 2022) is a large-scale dataset containing 747M image-text pairs as well as many other meta-attributes to increase the usability to train various models. MMC4 (Zhu et al., 2023) consists of 101.2 million documents with 571 million images interleaved in 43 billion English tokens. It covers a wide range of everyday topics such as cooking, travel, technology, and more. GRIT (Peng et al., 2023) is a large-scale dataset of Grounded Image-Text pairs that consists of approximately 91 million images, 115 million text spans, and 137 million associated bounding boxes. DataCamp (Gadre et al., 2023) is a participatory benchmark that focuses on dataset curation for large image-text datasets. It provides a new candidate pool of 12.8 billion image-text pairs. The dataset size in DataComp is a design choice and not predetermined.