SciMMIR: Benchmarking Scientific Multi-modal Information Retrieval

Anonymous ACL submission

Abstract

Multi-modal information retrieval (MMIR) is a rapidly evolving field where significant progress has been made through advanced representation learning and cross-modality alignment research, particularly in image-text pairing. However, current benchmarks for evaluating MMIR performance on image-text pairings overlook the scientific domain, which has a notable gap with the generic data since the caption of scientific charts and tables usually describes the analysis of experimental results or scientific principles in contrast to human activity or scenery depicted in generic images. To bridge this gap, we develop a **sci**entific domain-specific **MMIR** benchmark (**SciMMIR**) by leveraging open-access research paper corpora 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 from scientific documents. We further annotate the image-text pairs with a two-level subset-subcategory hierarchy to facilitate a more comprehensive evaluation of the baselines. We conduct zero-shot and fine-tuned evaluations on prominent multi-modal image-captioning and visual language models, such as CLIP, BLIP, and BLIP-2. Our findings offer critical insights for MMIR in the scientific domain, including the impact of pre-training and fine-tuning settings and the effects of different visual and textual encoders.

1 Introduction

Information retrieval (IR) systems are expected to provide a matched piece of information from an enormous, yet 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 of similarity matching approaches in a learned representation space, consequently supporting additional modalities such as images and audio, in addition to text (Karpukhin et al., 2020; Chen et al., 2020b; Koepke et al., 2022).

In scientific domains, offering users a fine-grained multi-modal retrieval service presents considerable practical significance. Although previous studies have evaluated the image-text retrieval task across a range of general topics on large-scale datasets such as Wikipedia (Young et al., 2014; Lin et al., 2014; Srinivasan et al., 2021; Luo et al., 2023), there is a notable research gap in comprehensively assessing MMIR models within the scientific domain, specifically. Integrating both in-domain and out-of-domain data in the pre-training phase significantly boosts the performance of visual language models (VLMs) on downstream tasks. However, most prior VLMs have focused exclusively on generic topic information of the mundane events in daily life, such as images depicting scenery and human activities, consequently overlooking data that is pertinent to scientific domains such as elements related to model architecture, illustrations of scientific principles, and results of experiments. Due to the substantial differences between the data distribution and characteristics between generic topic data and scientific data, many VLMs may not have an adequate ability to perform MMIR in the scientific domain. Additionally, past table-related work, such as table generation tasks, mainly focused on textual representations of tables while overlooking image-based representations of tabular data. This presents problems for human-computer interaction, as users may desire to input information in the form of screenshots and expect an interactive system to present results in a graphical format.

As shown in Figure 1, to address the identified research gap, we introduce **SciMMIR**, a **S**cientific **M**ulti-**M**odal **I**nformation **R**etrieval benchmark. SciMMIR is the first benchmark to comprehen-

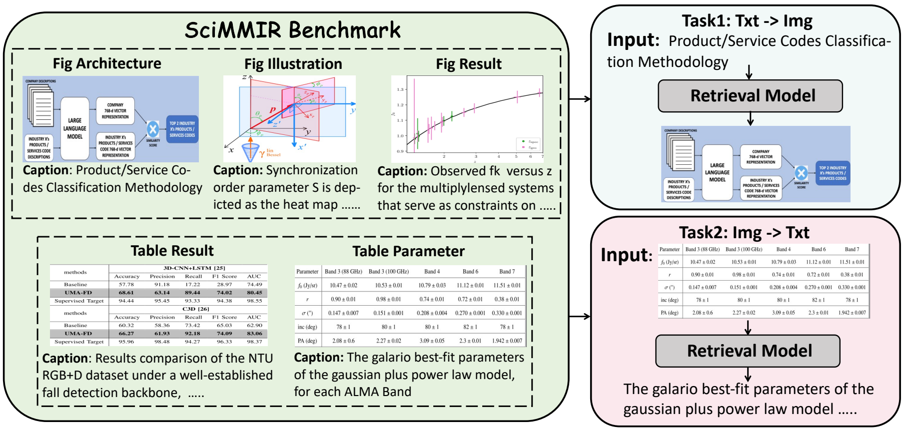

Figure 1: An illustration of the SciMMIR framework.

sively evaluate a model's MMIR ability in the scientific domain. To build our data collection, we retrieve the figures, tables (in form of image), and their associated captions, from scholarly documents available on arXiv, an open-access archival corpus, to construct image-text pairs. In order to comprehensively evaluate the cross-modality aligned representations learned by models, our SciMMIR benchmark defines the retrieval task as *bi-directional*, including searching the matched textual caption in candidate pool with a given image (`img`→`txt`), and finding corresponding figure or table image from a caption (`txt`→`img`).

The performance of VLMs across different types of data in the scientific domain is inconsistent, where a model may excel on data related to experimental results but demonstrate average performance with regards to image-caption pairs of model architectures. If an overall improvement is sought for the performance of VLMs, it may not yield a noticeable enhancement to its capabilities specifically regarding model architectures. Consequently, such improvements do not necessarily translate into effective boosts to a VLM's overall performance. Therefore, we annotate and categorise the image-text pairs into three figure-caption and two table-caption subcategories based on their distinctive described content (such as experimental results, model architectures, and scientific principles, etc.). Then we conduct ***fine-grained subset evaluation*** on subcategories in order to support targeted improvements to a model based on its performance in each subcategory, therefore potentially improving a model's capabilities by using high-quality data in a certain subcategory with a relative decrease in computational cost.

To explore the MMIR capabilities of our chosen image captioning models and VLMs in scientific domains, as well as different subcategories , we conduct extensive experiments in both zero-shot and fine-tuned settings across various subcategories. We present our key insights as follows:

1. We reveal that MMIR tasks in the scientific domain pose significant challenges for current VLMs, which usually do not demonstrate adequate performance in scientific domains. Furthermore, after fine-tuning VLMs with data specific to scientific domain, there is a marked performance improvement , underlining the effectiveness of domain-specific adaptation.
2. The results additionally suggest a distinction between tasks involving the figure and table subsets, with performance on the figure subset being more effectively improved by scientific data domain adaption, showing the generalisability of the visual encoders. In contrast, the performance of VLMs on the table subset is relatively weaker, likely due to image-text samples of tabular data seldom appearing during pre-training for the VLMs.
3. Regardless of parameter size, the BLIP-2 series of models generally perform better on SciMMIR than other pre-trained VLMs. This improved zero-shot capability may be the result of distinct pre-training tasks including image-text matching and image-text contrastive learning, rather than standard 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 is given in §5.

2 Related Work

General Information Retrieval. Information Retrieval is a fundamental task within NLP, and has recently been facilitated by dense representation learning (Reimers and Gurevych, 2019; Karpukhin et al., 2020). More recently, the desire for unified representations across tasks has become significant, with this line of research proposing to understand and evaluate task-agnostic representations 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 of IR models (Thakur et al., 2021). Through the subpar performance of general image-text models on SciMMIR, we evidence that scientific IR, especially when multi-modal, remains an out-of-domain (OOD) task despite advancements 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 linguistic representations. 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), with these small-scale retrieval datasets, in turn, becoming the standard evaluation approach for such systems. 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), and more recently, evaluating the unified cross-modal representations across diverse tasks has emerged as a prevalent trend (Wei et al., 2023).

Scientific Document Learning. Scientific information retrieval has received moderate attention 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) presents another major approach in evaluating vision-language systems, concerning in-depth visual grounding, rather than the use of distributional priors (Agrawal et al., 2018). It is in this area that work with a similar scope to ours in the scientific domain, such as PlotQA (Methani et al., 2020) and ChartQA (Masry et al., 2022), is seen. Our proposed SciMMIR benchmark distinguishes itself from these existing works by offering extensive coverage across annotations of figure and table subcategories, a larger dataset size, and the utilisation of the real-world data that is naturally paired and therefore not reliant on costly human annotation.

Subset	Subcategory	Number			Len (words)
		Train	Valid	Test	Caption
Figure	Result	296,191	9,676	9,488	52.89
	Illustration	46,098	1,504	1,536	38.44
	Architecture	13,135	447	467	27.27
Table	Result	126,999	4,254	4,229	27.23
	Parameter	15,856	552	543	17.10
	Total	498,279	16,433	16,263	43.19

Table 1: Statistics of the SciMMIR dataset.

3 Dataset Construction

Data Collection. We collect the PDF files from a 6 month period from arXiv via the official API.[1] We use an open-source tool (Clark and Divvala, 2016) to locate the non-textual elements (i.e., figures and tables) in the papers and then extract the corresponding caption texts. All tables and figures are stored in the form of images, and we remove the pairs that have empty captions. The aforementioned collection process results in the SciMMIR dataset that comprises 530K image-caption samples, with the average length of captions in the dataset being 43.19 words as shown in Table 1. The dataset is split into training, validation, and testing sets with $498,279$, $16,433$, and $16,263$ samples, respectively. As shown in Figure 2, the SciMMIR benchmark covers a multitude of disciplines. Amongst these, 10 disciplines account for more than 1%, such as Mathematics, Physics, and Computer Science. This attests to the diversity of our dataset and implies the presence of intricate scientific knowledge within.

Subset and Subcategory Structure. To better understand the performance of VLMs across various data types within the scientific domain, we define a hierarchical architecture with *two subsets* and

[1]We request data submitted between May and October 2023 from `https://info.arxiv.org/help/api`.

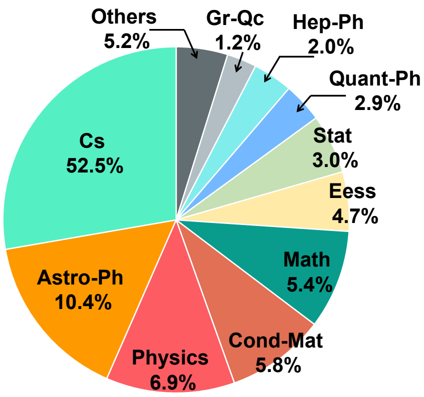

Figure 2: The ratio of different subject image-caption data in SciMMIR.

five subcategories for the SciMMIR benchmark. Initially, we divide the data into two subsets, Tables and Figures, as both representations have distinct data distributions. Tables contain ample textual information, whereas Figures predominantly utilize geometric shapes to elucidate scientific principles or reveal patterns within data. Furthermore, for tabular data, we further divide into two subcategories, Table-Parameter and Table-Result. This is performed as Table-Result data primarily serves to present experimental outcomes (i.e., numerical), whereas Table-Parameter data provides explanations of parameter meanings or specific numerical values (i.e., textual), and consequently both have different data type distributions. As for Figures, we consider those depicting experimental results, explaining model architectures, and illustrating various scientific theories to contain different elements of commonsense knowledge. Therefore, the performance of models on these distinct data types may vary, leading us to categorise them into three separate subcategories. The finer-grained categorisation is performed in accordance with Table 2.

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.

Data Annotation. In the process of data annotation, we use manually constructed key phrases to classify image-text sample pairs. Firstly, we obtain keywords by observing unique words that emerge in captions under different subcategories, thus conducting an initial categorisation of the data. Subsequently, to ensure the quality of our statistical analysis, we randomly select 2000 images from the test set and manually review the results of the keyword-based classification based on the criteria of whether the image within the image-caption pairs cater to the description of its subcategory. We then construct new keywords and remove low-quality ones by analysing which words in the caption result in misclassified examples. Finally, we iteratively construct a higher-quality list of keywords until the classification results of the extracted 2000 samples are deemed by manual evaluation as having achieved the optimal categorisation results. The subset and subcategory classification results are shown in Table 1, providing a structured and standardised basis for subsequent experiments.

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 multimodal information retrieval performance in scientific domains and distributions already learned by the models. To this end, we collect information about pre-training phase for baseline models in Table 3 and present additional details in Appendix A.

Image Captioning Models As our baselines, we present image-captioning models, including **CLIP-base** (Radford et al., 2021) and **BLIP-base** (Li et al., 2022), that have learned the pairing relationship between images and the corresponding text via a strong supervision signal. We evaluate these image captioning models trained on general domain datasets (such as images related to scenery and daily life events) in both zero-shot and fine-tuned settings to investigate the need for scientific domain adaption. We also introduce **BERT** (Devlin et al., 2018) as an alternative text encoder for captioning (denoted "+BERT" in the tables), where such ensemble baselines may reveal the influence of the text encoders.

Model	Pre-training Data: Domain	Pre-training Data: Number	Pre-training Task	Trainable & *Frozen Parameters: 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
LLaVA-V1.5-7B	LAION, CC, SBU, ShareGPT	392M	Language Modelling	0.3B	6.9B	0.02B

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

Visual Language Models. Additionally, we select large visual language models (VLMs) trained for multi-modal tasks such as VQA to examine their zero-shot and fine-tuning MMIR performance in scientific domain. The details regarding our chosen VLMs are presented in Appendix B.

4.2 Evaluation Protocol

Task Definition. The SciMMIR benchmark presents a bi-directional MMIR task:

- `txt→img`: The forward direction retrieval task, where given a corresponding text, the model must retrieve the correct image from a candidate set.
- `img→txt`: The inverse direction retrieval task, where given an image, the model must retrieve the relevant text from a candidate set.

For these two kinds of tasks, we all regard the samples of train, valid, and test data as candidates.

Given an image img_i and a text $text_j$, the relevance score R in the retrieval ranking is defined as the dot product between the visual and textual representations of img_i and $text_j$ by $R = E_{img_i} \cdot E_{text_j}$. In addition to assessing the model's performance on the overall test set (denoted "ALL" in the tables), we evaluate retrieval models in different subsets and subcategories to scrutinise their abilities. Specifically, we assess the model's performance on five fine-grained subcategories (shown in Table 2) of the test set, as well as the performance on the Figure and Table subsets as a whole.

Metrics. In this paper, we use the MRR and Hits@K metrics to assess the IR models' performance on the SciMMIR benchmark, which are further described in Appendix D.

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. We also provide evaluation of fine-tuned (FT) versions of the relatively small models (CLIP-base and BLIP-base) and a large VLM (BLIP2-FLAN-T5-XL) trained with our data. During fine-tuning, we employ standard contrastive learning (Chen et al., 2020a) to maximise the relevance score between positive text-image pairs and minimise the relevance score 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 (e.g., Figure-Result and Table-Parameter) of the training data to investigate the modeling abilities in a fine-grained manner.

5 Result Analysis

5.1 Overall Evaluation

Following the designed evaluation protocol, as shown in Table 4, we report the baseline performances in the universal set (ALL), Figure set, and Table set. In this subsection, we mainly discuss the

	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	8.13	13.48	7.94	13.34	9.29	15.41	8.99	15.29	5.29	8.82	5.41	8.65
	CLIP-base+BERT	2.47	5.01	3.11	5.85	2.99	6.09	3.80	7.10	1.19	2.42	1.44	2.85
	BLIP-base	6.14	11.30	6.18	11.71	6.80	12.59	6.89	13.21	4.59	8.22	4.47	8.15
	BLIP-base+BERT	**11.51**	**20.09**	**12.69**	**21.77**	**13.01**	**22.67**	**14.12**	**24.18**	**7.93**	**13.98**	**9.31**	**16.08**
	BLIP2-FLAN-T5-XL	4.44	7.74	2.27	4.48	4.93	8.66	2.57	5.02	3.23	5.48	1.51	3.13
ZS	CLIP-base	0.419	0.719	0.364	0.670	0.458	0.767	0.421	0.787	0.310	0.586	0.219	0.375
	BLIP-base	0.004	0.006	0.003	0.006	0.006	0.009	0.002	0.000	0.001	0.000	0.007	0.021
	BLIP2-FLAN-T5-XL	0.025	0.031	0.012	0.025	0.028	0.035	0.016	0.035	0.020	0.021	0.003	0.000
	BLIP2-FLAN-T5-XXL	0.053	0.105	0.004	0.000	0.059	0.104	0.004	0.000	0.040	0.105	0.003	0.000
	BLIP2-OPT-2.7B	0.052	0.111	0.015	0.031	0.035	0.060	0.013	0.027	0.093	0.230	0.020	0.042
	BLIP2-OPT-6.7B	0.002	0.006	0.002	0.000	0.003	0.008	0.002	0.000	0.002	0.000	0.002	0.000
	LLaVA-V1.5-7B	0.006	0.012	0.002	0.000	0.008	0.018	0.002	0.000	0.002	0.000	0.002	0.000
	mPLUG-Owl2-LLaMA2-7B	0.002	0.000	0.002	0.000	0.003	0.000	0.002	0.000	0.001	0.000	0.001	0.000
	Kosmos-2	0.008	0.018	0.002	0.000	0.011	0.025	0.002	0.000	0.000	0.000	0.001	0.000
	LLaMA-Adapter2-7B	0.040	0.061	0.002	0.000	0.056	0.085	0.002	0.000	0.001	0.000	0.004	0.000

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

results regarding the bi-directional retrieval tasks and the subset performance.

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 demonstrate superior performance in all settings of the 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. For tasks in either direction, many of the zero-shot large VLMs demonstrate insufficient performance, with the MRR and Hits@10 metrics, failing to surpass 0.23% in the ALL setting. It is worth mentioning that the CLIP-base model is well-trained since its zero-shot performance is better than all other large VLMs with superior parameter sizes.

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

Our SciMMIR benchmark demonstrates the shortcomings of VLMs in our SciMMIR task and provides extensive high-quality MMIR data for scientific domains that could be used for fine-tuning VLMs in order to improve performance on this domain. Additionally, our experiments show that retrieving visual tables is challenging and requires thoroughly mining the semantic connections between caption information and textual data within tables. For VLMs not adapted to the image-caption task in the scientific domain through pre-training (such as BLIP), fine-tuning with a vanilla pre-trained language model (such as BERT) can better establish connections between visual tables and captions due to captions for tables being a type of textual information rarely encountered by VLMs during their pre-training process.

5.2 Zero-Shot Analysis

To provide a more thorough analysis, we present the zero-shot performance of the baselines across different subcategories in Table 10 and Table 11 in Appendix F, where only the images or texts from the same subcategory are considered as candidates.

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 subcategories of the Table subset, all models, except CLIP-base, exhibit virtually no accuracy. In the Figure subset, the BLIP2-FLAN-T5 series of models show slightly better performance across all subcategories of the Figure subset. This may be attributed to the fact that the encoder part of text encoder-decoder architecture is better able to capture 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 the image-grounded text generation task of VLMs can enhance the model's performance in multi-modal retrieval for the forward direction, while the performance in the reverse direction is poorer.

5.3 Analysis on Fine-tuning Setting

Overall Analysis. As shown in Table 9 in Appendix E, we fine-tune the models using data of

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	9.77	16.92	9.84	15.42	10.01	15.30	9.35	14.97	9.16	15.37	8.90	15.34
	Fig-Architecture	5.60	8.35	6.11	8.14	2.61	4.95	2.95	5.01	2.50	4.02	2.35	4.18
	Fig-Illustration	8.58	12.85	8.82	13.28	6.76	11.72	7.08	11.78	5.69	9.20	5.46	8.96
	Fig-Result	9.24	15.42	9.76	14.99	8.58	14.19	8.86	14.26	8.79	14.10	9.05	14.79
	Table-Parameter	2.67	4.50	3.04	3.85	1.78	3.19	2.42	4.49	1.82	2.99	1.55	2.74
	Table-Result	3.12	5.78	3.31	5.35	1.91	3.91	2.33	4.49	2.58	4.26	1.48	2.80
CLIP-base+BERT	All	2.30	4.93	2.76	6.42	3.12	5.53	3.59	6.97	3.01	6.23	3.88	7.16
BLIP-base	All	5.11	10.06	5.53	10.28	5.35	10.09	5.64	10.16	7.11	13.10	7.15	13.82
	Fig-Architecture	0.04	0.00	0.06	0.21	0.02	0.00	0.03	0.07	0.03	0.06	0.02	0.01
	Fig-Illustration	0.04	0.00	0.09	0.00	0.26	0.52	0.45	0.91	0.08	0.16	0.09	0.14
	Fig Result	2.55	6.21	3.20	6.00	2.91	6.25	3.380	6.84	4.66	9.13	4.80	9.18
	Table-Parameter	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01	0.00
	Table-Result	0.12	0.21	0.01	0.00	0.01	0.00	0.03	0.07	0.05	0.07	0.06	0.09
BLIP-base+BERT	All	9.95	18.42	12.09	18.63	11.17	19.27	11.63	20.25	13.44	23.39	14.60	25.04
BLIP2-FLAN-T5-XL	All	6.75	11.34	4.06	8.56	5.99	10.41	3.16	6.44	4.69	8.27	2.41	4.64

Table 5: 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	5.40	9.01	5.52	8.82	4.45	7.37	4.55	7.37
	Fig-Architecture	1.22	2.06	1.34	2.34	1.35	2.58	1.47	2.95
	Fig-Illustration	1.42	2.70	1.79	3.14	1.93	2.95	2.60	4.42
	Fig-Result	2.71	4.49	2.53	4.52	2.19	4.05	2.30	4.79
	Table-Parameter	1.46	2.70	1.56	2.62	1.52	3.31	1.82	3.68
	Table-Result	4.28	7.26	1.28	2.29	3.77	6.63	0.87	1.29
CLIP-base+BERT	All	1.18	2.41	1.46	2.93	1.31	2.58	1.33	2.21
BLIP-base	All	4.77	8.42	4.54	8.23	3.16	6.63	3.99	7.55
	Fig-Architecture	0.01	0.00	0.03	0.02	0.01	0.00	0.02	0.00
	Fig-Illustration	0.00	0.00	0.01	0.00	0.01	0.00	0.02	0.00
	Fig-Result	0.70	1.32	0.65	1.16	0.32	1.29	0.56	0.74
	Table-Parameter	0.01	0.02	0.01	0.00	0.02	0.00	0.06	0.00
	Table-Result	0.92	1.80	0.92	1.82	0.83	0.74	0.52	1.10
BLIP-base+BERT	All	8.17	14.35	9.70	16.48	6.01	11.05	6.19	12.89
BLIP2-FLAN-T5-XL	All	3.11	5.29	1.33	2.90	4.22	6.99	3.00	4.97

Table 6: The results of fine-tuning models on Table subsets of our SciMMIR benchmark.

different categories and evaluate the performance regarding all samples in train, valid and test data 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 distributions among different subcategories.

The BLIP-base+BERT model performs the best across all fine-tuned settings, while the performance of the CLIP model decreases when its text encoder is replaced. Notably, merely fine-tuning the Q-Former parameters of BLIP2-FLAN-T5-XL to adapt the large VLM to the scientific domain did not yield as effective results as the smaller models. Consequently, there remains a need for efficiently fine-tuning small models to construct robust connections between the representations of the visual and textual modalities.

The Impact of Subcategory Training Data. As shown in Table 5 and Table 6, we report the result on testing samples of specific subcategories, for the sake of comprehensively investigating the impact of different subcategory training data.

For BLIP, the model's improvement on specific test subcategories generally aligns with the subcategories used for training, but its overall performance on the samples from other subcategories is poorer. This demonstrates the effectiveness of our annotation in accurately clustering data points, and the gaps among different subcategories.

As for CLIP, the models trained on different subcategories consistently perform 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. This suggests that solid pre-training can more effectively facilitate the model in adapting to the scientific domain, and further, it can potentially promote the model's learning of commonalities among different subcategories of data, thus enhancing its generalization capabilities across various subcategories.

Model	Testing Data	Fig-Architecture		Fig-Illustration		Fig-Result		Table-Result		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	12.85	12.72	16.62	18.22	69.57	67.22	0.84	1.65	0.13	0.19
	Fig Illustration	5.16	4.66	20.59	22.66	73.30	71.47	0.83	0.98	0.13	0.23
	Fig Results	3.80	3.62	13.01	14.25	81.48	80.15	1.48	1.64	0.22	0.34
	Table Results	0.12	0.15	0.24	0.70	4.16	4.97	85.68	84.29	9.81	9.89
	Table Parameters	0.29	0.35	0.53	1.34	5.08	9.61	73.44	72.19	20.64	16.50
ZS-CLIP-base	Fig Architecture	7.34	6.72	28.54	23.06	59.42	66.62	4.20	2.70	0.49	0.90
	Fig Illustration	3.99	3.68	30.56	23.44	61.74	71.04	3.40	1.47	0.31	0.36
	Fig Results	4.12	4.17	24.31	19.59	63.04	73.52	7.74	2.29	0.79	0.44
	Table Results	0.36	2.55	1.48	4.91	9.28	38.69	75.89	41.92	12.99	11.92
	Table Parameters	0.26	3.00	2.38	7.38	9.52	42.43	74.40	34.68	13.44	12.50

Table 7: The accuracy and error analysis of CLIP models on our SciMMIR benchmark.

The model trained on Figure-Results data demonstrates the best performance across the entire Figure subset. One reason could be that the Figure-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 highlights 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 SciMMIR, we substitute the text encoders in both BLIP-base and CLIP-base models with BERT-base. As shown in Table 9 in Appendix E, replacing the text encoder of BLIP with BERT results in a significant improvement, while that of CLIP experiences a decline. The reason for the performance change being opposite after replacing text encoder with BERT in both the CLIP and BLIP may be as follows:

CLIP. With the uniformity promise of contrastive learning (Wang and Isola, 2020), the textual and visual embeddings are well-aligned in an isotropic space in the pre-training phase of CLIP, which is demonstrated by the zero-shot setting experiments. However, replacing the text encoder with a highly anisotropic vanilla text encoder (e.g., 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 may help guide the replaced language model.

BLIP. On the other hand, in comparison to CLIP, BLIP uses BERT as its text encoder during the pre-training phase. This structural consistency contributes to the model's better adaptation to the scientific domain. However, the use of BERT may allow for the learning of a better representation of text to build an association between images and text, as tables contain a lot of text information.

5.5 Accuracy and Error Analysis

For better analysis of the performance, we conduct experiments on test data of different subcategories and calculate the ratio of all subcategories in the top 10 answers predicted by the fine-tuned CLIP and vanilla CLIP. Retrieval answers that have the same subcategory as the testing subcategories are regarded as correct, and vice versa.

As shown in Table 7, due to the larger volume of data for candidates labelled as Fig-Results and Table-Results (58.00% and 26.16%, calculated through Table 1, respectively), the models tend to predict samples from these categories as answers. When comparing zero-shot and fine-tuned models, it can be observed that fine-tuning leads to a decrease in the proportion of incorrect predictions across almost all categories.

Compared with zero-shot results, the fine-tuned models show the largest improvement in prediction accuracy on the Figure-Architecture and Figure-Result 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 categorise the images into fine-grained subcategories based on the characteristics of the figures and tables to facilitate a more comprehensive evaluation and analysis. Our evaluation of zero-shot and fine-tuned approaches, which we conduct on extensive baselines within various subsets and subcategories, offers valuable insights for future research.

Limitations

Due to computational resource constraints, we only fine-tune BLIP2-FLAN-T5-XL on our SciMMIR dataset and did not investigate the fine-tuning effects of other large VLMs on our benchmark. In this work, we find that BLIP+BERT could improve the model's ability in our benchmark, specifically for the Table subset. However, we do not design experiments to explore which kind of models would be better suited to the replacement of the textual encoder with BERT or other language models.

Ethics Statement

The dataset used in our research is constructed using publicly available data sources, ensuring that there are no privacy concerns or violations. We do not collect any personally identifiable information, and all data used in our research is obtained following legal and ethical standards In the stage of designing key words and human evaluation classification of image-text pair, we employed three graduate students experienced in natural language processing for human evaluation. We paid the graduate students about $13 per hour, well above the local average wage, and engaged in constructive discussions if they had concerns about the process.

Despite we try our best efforts to ensure data quality, given the sheer volume of data, we cannot guarantee that all results and content within the scientific domain dataset are accurate. This inherent limitation could potentially lead to models generating misleading or deceptive outputs in future use, necessitating further filtering in future work.

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.

Christopher Clark and Santosh Divvala. 2016. Pdffigures 2.0: Mining figures from research papers.

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*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023. Improved baselines with visual instruction tuning.

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 datasets 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. The 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. 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 with 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.

B Used Visual Language Models

- **BLIP-2** (Li et al., 2023) series models use a querying transformer module to address the modality gap. We choose 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.
- **LLaVA-V1.5-7B** (Liu et al., 2023) use two simple methods, namely, an MLP cross-modal connector incorporating academic task related data such as VQA to improve the ability of the LLaVA.
- **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) 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.

C Effects of Visual Encoder Resolution

In Table 4 for overall results, we compare the fine-tuned BLIP with the default image preprocessing dimensions of 384 and fine-tuned CLIP with the default image preprocessing dimensions of 224, where the results are relatively close. To make a fairer comparison, we decrease the image dimensions of BLIP-base model from 384 to 224 to be the same as CLIP-base to conduct SciMMIR evaluation, as described in Table 8.

It can be seen that the granularity of image processing has a significant impact on model performance. When using a lower preprocessing dimen-

Img Dim	Model	Training Dataset	txt→img MRR	txt→img Hits@10	img→txt MRR	img→txt Hits@10
224	BLIP-base	ALL	0.958	2.034	1.138	2.294
		Fig Architecture	0.002	0.000	0.006	0.000
		Fig Illustration	0.036	0.024	0.011	0.000
		Fig Result	0.167	0.260	0.115	0.213
		Table Result	0.408	0.757	0.368	0.686
		Table Parameter	0.011	0.024	0.009	0.000
224	BLIP-base+BERT	ALL	1.614	3.334	2.102	4.375
384	BLIP-base	ALL	6.14	11.3	6.18	11.71
		Fig Architecture	0.02	0.04	0.02	0.02
		Fig Illustration	0.07	0.14	0.1	0.17
		Fig Result	3.26	6.48	3.4	6.5
		Table Result	0.3	0.54	0.3	0.57
		Table Parameter	0.01	0.01	0.01	0
384	BLIP-base+BERT	ALL	11.51	20.09	12.69	21.77

Table 8: The averaged results of fine-tuning BLIP with different preprocessing image dimensions on *ALL* testing candidates of our SciMMIR benchmark.

sion, 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 better establish the connection between image and text representations.

D MRR and Hit@K

- **MRR** stands for Mean Reciprocal Rank, and is calculated as 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 measurements.

E Fine-tuning Analysis

The effect of text-image matching task. As shown in the Table 9, the BLIP-2 series of models outperform other large VLMs in both Figure and Table subcategories, especially in the forward direction task. We believe that this is because BLIP-2 incorporates the text-image matching task and the image-grounded text generation task during its pre-training process to better align textual and visual information. 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 for multi-modal retrieval still requires a primary focus on the text-image matching task.

Model	Training Dataset	txt→img MRR	txt→img Hits@10	img→txt MRR	img→txt Hits@10
CLIP-base	ALL	8.13	13.48	7.94	13.34
	Fig-Architecture	2.23	3.67	2.22	3.86
	Fig-Illustration	4.64	7.64	4.66	7.69
	Fig-Result	6.98	11.31	7.13	11.74
	Table-Parameter	1.74	2.99	1.68	2.94
	Table-Result	3.01	5.13	1.54	2.85
CLIP-base+BERT	ALL	2.47	5.01	3.11	5.85
BLIP-base	ALL	6.14	11.30	6.18	11.71
	Fig-Architecture	0.02	0.04	0.02	0.02
	Fig-Illustration	0.07	0.14	0.10	0.17
	Fig-Result	3.26	6.48	3.40	6.50
	Table-Parameter	0.01	0.01	0.01	0.00
	Table-Result	0.30	0.54	0.30	0.57
BLIP-base+BERT	ALL	11.51	20.09	12.69	21.77
BLIP2-FLAN-T5-XL	All	4.44	7.74	2.27	4.48

Table 9: The results of fine-tuning models that are trained on different types of training data.

F Zero-shot Analysis

CLIP-base and BLIP-base. As shown in the Table 10 and Table 11, 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 effec-

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	1.351	1.927	1.074	2.141	0.750	1.237	0.458	0.716	0.373	0.643	0.386	0.738
BLIP-base	0.003	0.000	0.001	0.000	0.003	0.000	0.002	0.000	0.006	0.011	0.002	0.000
BLIP2-FLAN-T5-XL	0.010	0.000	0.003	0.000	0.010	0.000	0.004	0.000	0.032	0.042	0.019	0.042
BLIP2-FLAN-T5-XLL	0.056	0.214	0.003	0.000	0.037	0.065	0.005	0.000	0.062	0.105	0.004	0.000
BLIP2-OPT-2.7B	0.130	0.214	0.005	0.000	0.033	0.130	0.006	0.000	0.031	0.042	0.014	0.032
BLIP2-OPT-6.7B	0.001	0.000	0.001	0.000	0.009	0.065	0.001	0.000	0.002	0.000	0.002	0.000
LLaVA-V1.5-7B	0.003	0.000	0.004	0.000	0.003	0.000	0.004	0.000	0.009	0.021	0.002	0.000
Kosmos-2	0.123	0.428	0.008	0.000	0.011	0.000	0.004	0.000	0.006	0.011	0.002	0.000
mPLUG-Owl2-LLaMA2-7B	0.022	0.000	0.003	0.000	0.302	0.521	0.003	0.000	0.019	0.021	0.002	0.000
LLaMA-Adapter2-7B	0.001	0.000	0.001	0.000	0.008	0.000	0.002	0.000	0.002	0.000	0.002	0.000

Table 10: 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	0.281	0.544	0.177	0.284	0.545	0.921	0.558	1.105
BLIP-base	0.001	0.000	0.007	0.024	0.000	0.000	0.003	0.000
BLIP2-FLAN-T5-XL	0.021	0.024	0.003	0.000	0.010	0.000	0.005	0.000
BLIP2-FLAN-T5-XLL	0.041	0.095	0.003	0.000	0.030	0.184	0.003	0.000
BLIP2-OPT-2.7B	0.076	0.213	0.010	0.024	0.228	0.368	0.101	0.184
BLIP2-OPT-6.7B	0.002	0.000	0.002	0.000	0.001	0.000	0.002	0.000
LLaVA-V1.5-7B	0.002	0.000	0.002	0.000	0.003	0.000	0.004	0.000
Kosmos-2	0.000	0.000	0.001	0.000	0.000	0.000	0.003	0.000
mPLUG-Owl2-LLaMA2-7B	0.001	0.000	0.004	0.000	0.002	0.000	0.005	0.000
LLaMA-Adapter2-7B	0.001	0.000	0.001	0.000	0.001	0.000	0.001	0.000

Table 11: The zero-shot results of multi-modal models on Table subsets of our SciMMIR benchmark datasets.

tive learning on the scientific domain multi-modal data. These models have strong MMIR abilities for generic topic data, such as BLIP achieving an IR@1 of 86.7% on the Flicker dataset in the zero-shot setting, whilst BLIP does not surpass 0.05% (MMR metric). This further demonstrates the challenges presented for MMIR in scientific domains.