SciOL and MuLMS-Img: Introducing A Large-Scale Multimodal Scientific Dataset and Models for Image-Text Tasks in the Scientific Domain

Tim Tarsi[1] Heike Adel[2] Jan Hendrik Metzen[1] Dan Zhang[1] Matteo Finco[4] Annemarie Friedrich[3]
[1]Bosch Center for Artificial Intelligence, Renningen, Germany
[2]Hochschule der Medien, Stuttgart, Germany [3]University of Augsburg, Germany
[4]Robert Bosch GmbH, Renningen, Germany
tim.tarsi@gmail.com {janhendrik.metzen|dan.zhang2|matteo.finco}@de.bosch.com
heike.adel@gmail.com annemarie.friedrich@informatik.uni-augsburg.de

Abstract

In scientific publications, a substantial part of the information is expressed via figures containing images and diagrams. Hence, the retrieval of relevant figures given a natural language query is an important real-world task. However, due to the lack of training and evaluation data, most existing approaches are either limited to one modality or focus on non-scientific domains, making their application to scientific publications challenging.

In this paper, we address this gap by introducing two novel datasets: (1) SciOL, the largest openly-licensed pretraining corpus for multimodal models in the scientific domain, covering multiple sciences including materials science, physics, and computer science, and (2) MuLMS-Img, a high-quality dataset in the materials science domain, manually annotated for various image-text tasks. Our experiments show that pre-training large-scale vision-language models on SciOL increases performance considerably across a broad variety of image-text tasks including figure type classification, optical character recognition, captioning, and figure retrieval. Using MuLMS-Img, we show that integrating text-based features extracted via a fine-tuned model for a specific domain can boost cross-modal scientific figure retrieval performance by up to 50%.

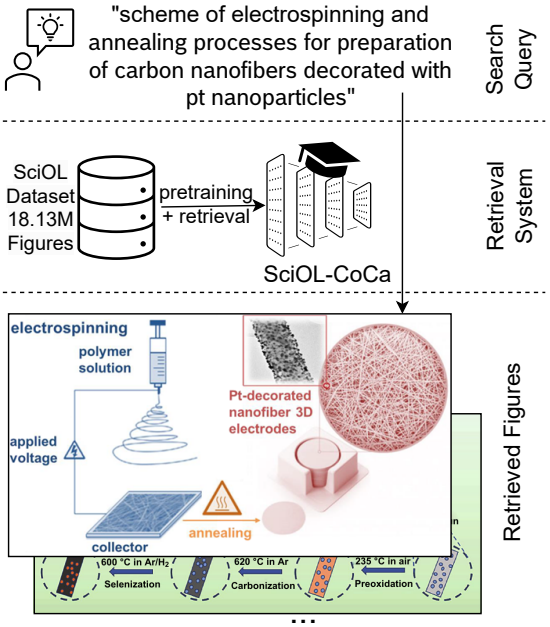


Figure 1: Real-world application of SciOL-CoCa: Scientific figure retrieval for the materials science domain.

1. Introduction

In research and development alike, it is crucial to be aware of state-of-the-art scientific results, which are usually published in the form of scientific articles. The number of potentially relevant publications is typically very large and grows exponentially [5], and targeted automatic information retrieval systems could be of great value to solving real-world problems. Most systems to date, however, are limited to textual retrieval. Yet, figures play a central role in scientific publications, often conveying important information including experimental outcomes. Image retrieval systems are predominantly trained on images from non-scientific domains, which are distinctly different from those in scientific fields [70]. While non-scientific images often show natural scenes or objects, scientific images might feature curves in plots or specific data visualizations. Existing scientific image retrieval models [16, 41, 70] are limited to the biomedical and life science domains. They leverage image-text models designed for natural images [47]. These

systems have two shortcomings: (1) they are biased towards the training data domain and (2) they rely on the model's ability to implicitly understand text within figures at pixel level and disregard additional information such as captions.

In this paper, we address this gap by introducing two novel datasets. We first present SciOL (*Openly-Licensed Scientific Publications*), a large-scale automatically derived dataset consisting of scientific publications including their image-caption pairs, which can, i.a., be used as pre-training data for various downstream tasks such as figure retrieval or captioning. In contrast to existing similar datasets [23, 70], SciOL covers a broad variety of scientific fields, including materials science, mathematics, physics, computer science and economics. The included papers all have permissive licenses, ensuring a broad applicability of our dataset.

Second, we present the *Multi Layer Materials Science Image Corpus (MuLMS-Img)*, providing high-quality manual annotations for figure type classification and text recognition in scientific figures. It also contains queries written by a domain expert, enabling us to perform figure retrieval experiments and to evaluate our models trained on SciOL in a real-world setting (see Figure 1).

In our experiments, we find that pre-training state-of-the-art models on SciOL improves performance across vision and vision-language tasks by a large margin. Our results also show that integrating figure captions and the output of optical character recognition (OCR) as features leads to a performance boost of around 50% for cross-modal scientific figure retrieval in a specific domain. This indicates the importance of our annotations for different multi-modal tasks, and shows how our pre-trained model can be adapted to further domains with limited effort. To sum up, we make the following contributions:

- We build and release SciOL, a large-scale openly-licensed dataset of English scientifc publications converted into semi-structured data. It consists of over 18 billion tokens and over 18M image-text pairs extracted from 2.75M publications, making it, to our knowledge, the largest collection of openly licensed text- and image-text data.
- We present MuLMS-Img, a high-quality dataset from the materials science domain, annotated for a variety of text-image tasks, i.e., figure type classification, OCR and text role labeling, and figure retrieval.
- We train different models for scientific figure captioning and retrieval, tailoring them to the scientific domain and real-world applications via pre-training on SciOL and investigating the effects of combining the results of different text-image tasks from MuLMS-Img.
- We open-source the SciOL metadata-catalog with information about source and contents of the publications and the semi-structured pre-training corpus.[1]

[1] https://github.com/boschresearch/sciol-wacv-2024

2. Related Work

Multimodal and Domain-Specific Pretraining. Vision language pretraining (VLP) aims at learning universal and transferable image and text features that generalize well on a variety of downstream applications [67, 1, 33] through multimodal pre-training tasks, such as image-text retrieval [59, 72, 7, 69, 30, 27, 35, 3, 47, 34], caption generation [74, 12, 66, 64], and text-to-image generation [49, 51, 68]. Inspired by the great success in natural language processing, transformer-based architectures [62] have also gained popularity in computer vision [15]. Multimodal feature fusion has shown to capture fine-grained dependencies between visual and text features [7, 30, 35, 3, 34, 33], e.g., through inter- and intra-modality attention. Another recent trend uses unified architectures to jointly optimize multiple objectives [34, 67, 1, 33]. To the best of our knowledge, our work is the first including an autoregressive captioning task [67] in addition to generating visual and textual representations into the pre-training of a scientific vision-language model. Recent methods for VLP in the scientific domain use medium to large-scale scientific pretraining datasets crawled from indexing services [19, 23, 16, 70]. Our method is most similar to [65, 41, 24, 16, 70], who pretrain models based on the CLIP architecture [47] on larger-scale scientific datasets with an image-text matching objective. However, these models focus on either the biomedical or the computer science domain only. In contrast, our work covers a broader range of scientific topics.

Scientific Figure Retrieval and Captioning. Similar to the natural image domain, scientific figure retrieval is based on aligning image and text embeddings in a joint feature space by encoding the text and pixel information with uni-modal networks [72, 65, 41, 24, 16, 70]. Recent work on captioning and figure question answering propose taking into account the special structure of scientific figures. These models aim at generating a structured representation for each image, which are the basis for further processing using heuristic [6, 2] or neural networks [57, 43]. While this works well for charts, it does less so for other types of scientific figures such as illustrations or photographs. Our work combines the strengths of these approaches: we leverage powerful image-text embeddings from a pretrained large-scale vision-language model. However, because text plays a crucial role in scientific figures, we also explicitly detect and extract the text within figures, enhancing image representations by embeddings of this extracted text.

Scientific Datasets. Application-oriented datasets span a wide range of modalities and tasks, including vision (e.g., classification [63], detection [11]), language (e.g., concept extraction [18]), and their intersection (e.g., visual question answering [29, 42, 6, 32], image captioning [43, 50], and data extraction [11]). These datasets are mostly manually curated and annotated to develop and evaluate task-specific

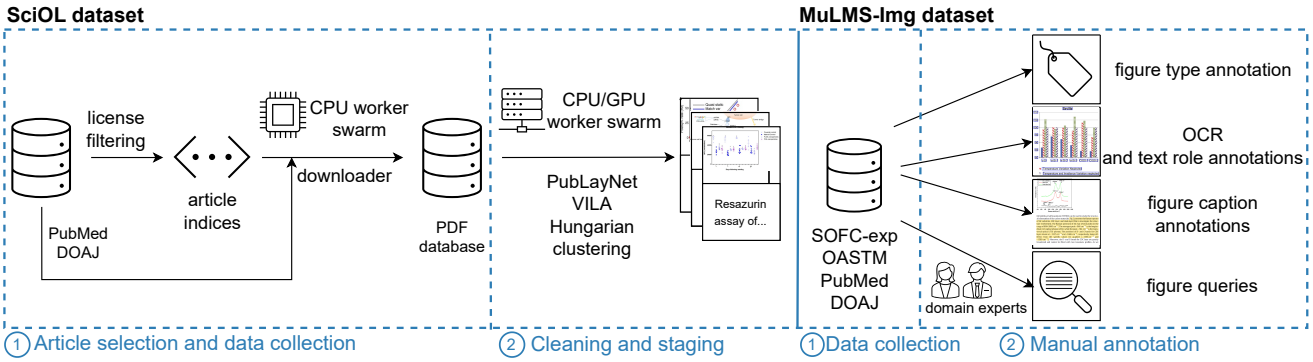


Figure 2: Construction of our datasets: left: SciOL, right: MuLMS-Img.

models. To the best of our knowledge, MuLMS-Img, which we present in this work, is the first image-text dataset in the materials science domain.

Existing scientific pre-training datasets for VLP focus on particular domains, e.g., the (bio-)medical domain [70] or computer science [23]. In contrast, we collect publications from various scientific domains, such as computer science, physics, engineering, and mathematics. We ensure that all content is openly licensed and publish the full text and metadata alongside the image-caption pairs.

3. New Image-Text Datasets for Science

In this section, we describe the collection, annotation and corpus statistics of each dataset, and analyze the quality of our new corpora. The construction is illustrated in Figure 2.

3.1. SciOL: Scientific Openly-Licensed Dataset

We first present SciOL, the largest scientific multimodal and permissively licensed image-text corpus to date.

3.1.1 Data Collection and Integration

To select journals and articles, we use the indexing services of PubMed[2] and DOAJ[3]. We exclusively select articles licensed under CC-0, CC-BY 2.0, CC-BY 3.0, CC-BY 4.0, MIT and Apache2.0, to ensure the availability for both academic and industrial research. While the PubMed index already includes license information, the DOAJ endpoint does so only for journals. Therefore, we first filter journals by the specified licensing terms and select articles from these journals based on the ISSN or eISSN. As journals might not publish under an exclusive license, we treat the selected articles only as candidates, where the licensing needs to be confirmed either when downloading the articles or during cleaning. We keep only English articles and perform deduplication.

[2]We use the PMC Open Access - Commercial Use Subset [44] as index

[3]We access DOAJ through the OAI-PMH endpoint [45].

PubMed offers the download of article files through the individual download endpoint of the FTP service [44]. DOAJ only indexes publications but does not distribute them. We found that the provided URL leads to the article PDF only for 20% of our selected articles. In the remaining cases, the URLs direct to the webpages of the journal where the PDF resource is linked. To retrieve the PDF data, we either extract its web address from the response header if present, or search for keywords in the display text indicating anchors that contain the URL referring to the PDF data. However, this process is noisy and generates multiple candidates for each article, as the webpage might contain URLs for additional file formats or other files. We download all files and filter out false positives.

3.1.2 Cleaning and Staging

We next prepare the data for processing in the context of VLP tasks (Section 4). We first convert the PDF files into semi-structured data, i.e., we extract and structure text elements and match figures and corresponding captions. The PubMed data provides images and text in a semi-structured XML format. We clean the text from typesetting commands and extract the captions associated with each image. For DOAJ-indexed files, we clean the raw downloads by removing all non-PDF files and filtering by license. We then convert the PDFs into a semi-structured format. We apply PubLayNet [73] with the layoutparser [55] framework for layout detection and VILA [54] for text extraction and word-level classification into categories such as body text, section header or caption. We then group the text into paragraphs using the DBScan clustering algorithm [17] keeping the linear text ordering the same as in the original.

Next, we extract figures and match them to the corresponding captions. Since captions might be positioned above or below figures, we find the optimal matching with the Hungarian algorithm [31]. We set the cost to the minimum of the absolute pixel distance between the top of the

	Domain	Avg. # Tokens	# Figures
LAION 2B	web images	9.9	2300M
SciCap	computer science	41.3	0.42M
PMC15M	biomedicine	110.0	15.28M
SciOL (ours)	science	75.8	18.13M

Table 1: Corpus statistics for various multimodal datasets. Numbers for PMC15M are from [70]. For the other datasets we use the NLTK punkt tokenizer [4] and estimate the avg. # tokens for LAION 2B from 10% of all samples. LAION 2B is constructed from the Common Crawl [53].

caption and the bottom of the figure and the bottom of the caption and the top of the figure. As this results in a bipartite matching, false positive figure detections are typically discarded. In addition, we provide the extracted *body text* of the publications, i.e., any text that does not belong to the title, abstract, bibliography or captions.

3.1.3 Computational Effort

In the data collection step (Section 3.1.1), we process about 25 TB of data over a period of five days, distributing the load on 20 single CPU workers. For cleaning and staging (Section 3.1.2), we distribute the processing on both multi-CPU only and GPU workers. The layout detection and text extraction and classification are performed on V100 GPUs. For the clustering, rendering and matching we employ CPU workers. Parsing all files takes approximately 150 single-worker-days, with the inference of VILA and rendering of the PDFs being the major bottleneck.

3.1.4 Dataset Statistics

Table 1 compares the statistics of SciOL to prior multi-modal datasets. With over 18M caption-image pairs, SciOL is the biggest multi-modal corpus in the scientific domain, and also the largest collection of permissively licensed image-text pairs. While PMC15M also contains around 15M figures, it has not been filtered for open licenses and is limited to the biomedical domain. According to Table 1, scientific figures have much longer captions than web images, making the modeling more challenging, which is in line with prior work [23, 70]. The average caption length of SciOL is between that of SciCap (Computer Science and Machine Learning) and that of PMC15M (biomedical and life sciences), which is expected as it includes articles of several scientific domains. In total, SciOL contains more than 18 billion tokens for language modeling which makes it roughly 25 times larger than the permissively licensed arXiv abstracts corpora [61] and about 5 times larger than the PubMed abstracts corpora [20] (Table S1).

	Body Text				Captions
Approach	S↓	I↓	D↓	WER↓	WER↓
Ours	**6.9**	**14.5**	**8.5**	**29.9**	**36.1**
pdffigures2.0+pdfplumber	38.5	15.2	33.3	87.0	64.1

Table 2: Extraction quality in terms of substitutions (S), insertions (I), deletions (D) and word error rate (WER) in %.

3.1.5 Corpus Quality Analysis

To evaluate the quality of data conversion and extraction, we make use of the PDF and XML files provided by PubMed. The latter also contains the plain text, hence, we can use them as the ground truth. We randomly sample 1000 publications and process the PDF files with our pipeline. As evaluation metric, we use the word error rate (WER), i.e., the sum of the number of substitutions, insertions and deletions needed to convert the extracted text into the ground-truth text, normalized by the number of words per document.

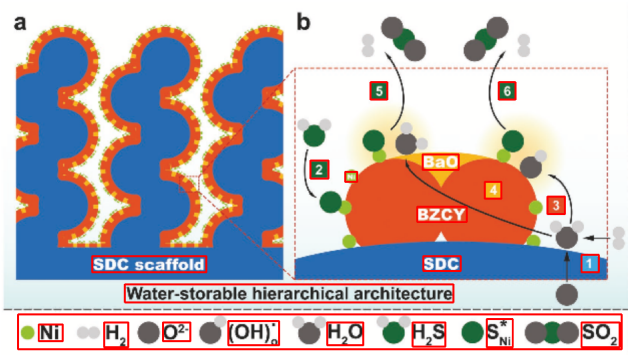
We evaluate the extraction of captions and body text including the abstract separately. If a caption has not been retrieved, the WER is set to 100. The individual scores are averaged on document level. As out-of-the-box baseline, we extract captions and images with pdffigures2.0 [9] and body text with pdfplumber [56]. Table 2 presents the results, showing that our tailored approach is effective. According to our qualitative analysis, in contrast to our approach, the baseline seems to have issues with multi-column layouts.

3.2. MuLMS-Img Dataset

The Multi-Layer Materials Science (MuLMS) corpus [52] is a dataset of 50 scientific publications in the materials science domain annotated for various natural language processing tasks. MuLMS-Img extends this dataset by providing annotations for various image-text tasks.

3.2.1 Data Selection and Annotation Process

The 50 articles included in MuLMS stem from the SOFC-Exp corpus [18], the OA-STM corpus [28], PubMED and DOAJ, and are all licensed under CC-BY. Image-level annotations, such as class labels, OCR and text role annotations, are created using the *labelimg* [60] tool. The two annotators, who are part of a professional in-house team specialising in image annotation services, do not have a background in materials science, but for the type of annotations they created, no deeper interpretation of the diagrams was necessary. The retrieval queries are written by two experts from the materials science domain, one graduate student of the materials science and one materials science researcher with a degree in environmental engineering. During annotation, we constantly checked the data quality based on our guidelines and discussed corner cases with the annotators.



Query 1: Graphical representation of the steps involved in water-induced sulfur removal from sofc anode ni nanoparticles on bzcyn in presence of amorphous bao.
Query 2: Sulfur removal schematized mechanism induced by water in sofc anode with ni nanoparticles and amorphous bao.

Figure 3: Image (source: [58]) from MuLMS-Img annotated with bounding boxes and text queries for retrieval.

3.2.2 Tasks and Annotation Guidelines

In materials science publications, information about experiments and their outcome is often spread across figures, captions, and body text. For an effective figure captioning and retrieval, it is essential to combine this information. To develop computational models for this challenge, we introduce the following subtasks (partially illustrated in Figure S2) and provide human annotations for them:

Figure Type Classification constitutes a multi-class classification task of identifying the type of a figure, e.g., chart types such as *bar plot*, *photograph* or *illustration*. These annotations could be used to restrict retrieval on particular figure types. Each figure is labeled with a class label from a taxonomy of nine chart types, photographs and illustrations (for a full list see Supplementary S3.2.2). We build on existing taxonomies, such as UB-PMC [11], but merge horizontal and vertical classes, e.g., for bar charts, and add two classes (*photography* and *illustration*).

Optical Character Recognition (OCR) and Role Labeling requires bounding-box detection and transcription of the text within the bounding box, plus identifying the role of the content in the figure, e.g., ticks, legends, or axis labels (for a full list see Supplementary S3.2.3). Typesetting, following the LaTeX syntax, is used for mathematical expressions, such as subscripts or superscipts.

Figure Retrieval is based on brief, *search-style* textual queries. Our aim is to create real-world search queries that might be used in a retrieval system, where the style typically deviates from the descriptive and wordy nature of captions. For selected figures from MuLMS-Img, a domain expert (a materials science graduate student) provides textual queries.

Annotation Type	Count
Figure type	1,075
OCR annotations	13,701
Figure retrieval queries	78

Table 3: Corpus statistics for MuLMS-Img.

3.2.3 Statistics of Dataset

Table 3 shows the number of annotations for the MuLMS-Img dataset. The distribution of figure type annotations is highly imbalanced with the two majority classes accounting for $\geq 51\%$ of all samples (Supplementary S3.3). Each image contains 12.7 OCR annotations on average. The location distribution of the bounding box annotations is skewed towards the left side and bottom of the images (Supplementary S3.3). Text within these annotations have an average length of 10.2 characters. MuLMS-Img contains a total of 78 figure retrieval queries. Due to the search-style writing, queries only consist on average of 10.5 tokens, which is much lower than the average token count for scientific figure captions (compare Table 1).

4. SciOL-CoCa: A Large-Scale Vision-Language Model for the Scientific Domain

In this section, we illustrate how a model pretrained on SciOL can be used for downstream tasks using MuLMS-Img as benchmark dataset. Figure 4 provides an overview of the training and application process.

4.1. Model Architecture

We use the CoCa architecture developed for building image-text representations and image captioning for non-scientific images [67] and propse the following modifications for the scientific domain.

Vision encoder: Following the CoCa-ViT-B32 [26] implementation, we employ a Vision Transformer [14] with 12 layers and learnable positional encodings as the image encoder. Scientific figures, however, come with a special challenge compared to images from other domains: they often contain fine-grained contents and include text within the figures. To be able to also attend to such information, we select an input resolution of 300×300 pixels and set the patch size to 16×16 pixels (compared to 224×224 and 32×32 for CoCa-ViT-B32 [26]).
Text encoder: Similar to Yu et al. [67], we use a Transformer [13] as unimodal text encoder. In Section 3.1.4, we have shown that captions in scientific text tend to be rather long. To model this, we increase the maximum input sequence length from 77 to 256 tokens allowing us to model 90% of the captions in our corpus without information loss.
Text decoder: The causal text decoder follows the archi-

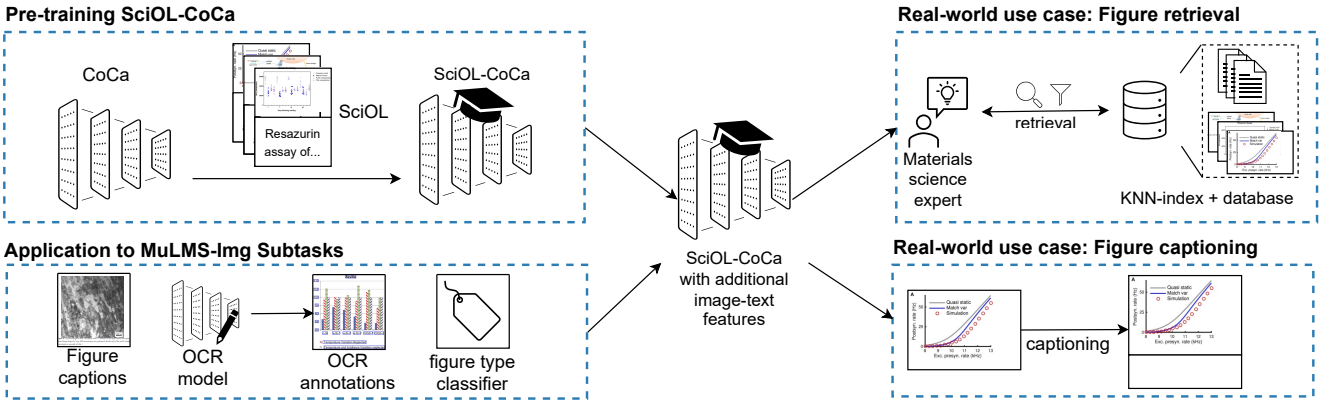


Figure 4: SciOL-CoCa: model pre-training, fine-tuning and application to real-world use cases.

tecture of the CoCa-ViT-B32 model [67]. It uses cross-attention to combine the features from the two unimodal encoders. Similar as for the text encoder, we change the maximal output sequence length from 77 to 256 tokens.

4.2. Pretraining on SciOL

We sample 85% of the SciOL caption-image pairs at random to construct a pretraining dataset. The remaining 15% of the data are used for validation (5%) and testing (10%). The dataset split will be made available for reproducibility. Detailed training settings can be found in Supplementary S5.

Following the recent trend of using multitask learning during pretraining [34, 33, 67], we leverage SciOL for image-to-caption generation, and representation learning pretraining. We jointly optimize two objective functions during training as proposed by Yu et al. [67]: an image-caption generation loss and a contrastive learning loss for representation learning. Similar to Li et. al [34] we employ momentum encoders [21] for the visual and text encoder to increase the number of negative samples for contrastive learning. Details on the joint objective can be found in Supplementary S4. To reduce the memory footprint and required computing resources during pre-training, we follow [37] and randomly mask 50% of the visual tokens during all training epochs except for the final one where we disable masking. Ablations on the effect of this final unmasked finetuning can be found in Supplementary S6.1.

4.3. Application to Downstream Tasks

We apply SciOL-CoCa to several downstream tasks and real-world use cases: figure type classification, captioning and retrieval. For figure caption generation and zero-shot figure type classification, we encode the figure with the vision encoder of SciOL-CoCa. For the former, we then generate the most-likely caption using its text decoder. For the latter, we compute the cosine similarity between the image embedding and the text embeddings of the class labels and select the label with the highest matching score as the predicted class. SciOL-CoCa is also the backbone for our figure retrieval system, which we describe next.

5. Application to Figure Retrieval (MuLMS-Img)

For extracting additional text-based features from scientific figures and text, we fine-tune models for the MuLMS-Img tasks, using a combination of MuLMS-Img and 2022 CHART-Infographics UB-Unitec PMC Dataset (UBPMC) [11] shared task data. The latter contains image annotations such as class labels, bounding box and text annotations for the related domain of biomedical chart figures. For ablations on the training data and setup see Supplementary S6.

For extracting text from images, we propose Sci-TrOCR, a two-stage model for text detection and text recognition. First, for identifying text within the image, we use a faster RCNN [48] with feature pyramids [39] as bounding box regressor for textual region and initialize the feature extractor with weights pretrained on MS-COCO [40]. In the second stage, we use TrOCR-Base [36] to transcribe text from the extracted regions and initialize with pretrained weights on the SROIE dataset [25].

Figure retrieval requires a representation of the textual query as well as a representation of the images in our corpus. We pre-prend a special classification token `[CLS]` to the textual query and embed it with SciOL-CoCa. We use the embedding of this token as query embedding. For the figure embeddings, we encode the figures with the vision encoder of SciOL-CoCa. For ranking figures with regard to a textual query, we use cosine similarity between the respective embeddings.

To test our hypothesis that it may be beneficial to leverage textual information explicitly extracted from the images, we combine the pre-trained SciOL-CoCa model from

	UBPMC		MuLMS-Img	
	F1-mac.↑	F1-mic.↑	F1-mac.↑	F1-mic.↑
BioMed-CLIP [70]	32.1	34.7	17.7	26.1
CoCa-ViT-B32 [26]	36.5	44.0	33.9	54.6
SciOL-CoCa	**39.4**	**46.0**	**39.8**	**61.4**

Table 4: Zero-shot figure type classification results in %.

Section 4 with the text-based features described above for our real-world use case of figure retrieval. We aggregate the image representation obtained by the vision encoder of SciOL-CoCa with a vector representing the caption and a vector representing the OCR-extracted text, each L2-normalized to unit scale.

- **Caption.** We encode the caption of each figure with the text encoder of our SciOL-CoCa model and use the latent representation of the CLS-token as feature vector.
- **OCR.** We use our OCR network to extract text from each image. The text is aggregated into a single comma-separated query and encoded with SciOL-CoCa as OCR feature vector.

6. Experiments and Results

To demonstrate the effectiveness of large-scale pretraining on our proposed SciOL dataset, we evaluate our models on multiple downsteam tasks in the scientific domain.

6.1. Figure Type Classification Experiments

We evaluate our model on the MuLMS-Img figure type annotations and on the UBPMC [11] dataset in a zero-shot setting, i.e., we use the model to encode both the image and the string representations corresponding to the names of the figure types and rank the latter by cosine similarity w.r.t. the image embedding.

UBPMC consists of chart images with fine grained figure type annotations, e.g., "horizontal interval chart" or "Manhattan plot." Following Davila et. al [11], we use F1-macro as our primary evaluation metric and report the micro scores in addition (further results see Supplementary S6.3).

Results. Table 4 shows the large positive effect of pretraining on our SciOL dataset (SciOL-Coca vs. CoCa-ViT-B32). In addition, we compare our results to BioMed-CLIP [70], a CLIP-based model pre-trained on PubMed. BioMed-CLIP performs considerably worse, especially on MuLMS-Img. We find that SciOL-CoCa performs better or on par across all classes, with the highest increase in class-wise F1 score for "micrograph/photograph" ($\Delta F1 = 40.8$) and "area chart" ($\Delta F1 = 66.7$).

6.2. Figure Captioning Experiments

We evaluate SciOL-Coca on the downstream task of figure captioning, ensuring that the evaluation data does not overlap with SciOL. As SciCap [23] and ImageCLEF [50], which are collected from PubMed and ArXiv, might contain overlaps, we resort to the HCI alt-text dataset [8] for our evaluation. It consists of 3386 scientific figures with alt-text descriptions extracted from publications on Human-Computer Interaction and accessibility. To measure the similarity between the generated and ground-truth descriptions, we use BERT-Score [71] and ROUGE [38].

	HCI-alt-text charts		SciOL test	
	Rouge↑	BERTScore↑	Rouge↑	BERTScore↑
CoCa-ViT-B32 [26]	7.7	78.2	8.7	78.4
SciOL-CoCa	**14.4**	**81.3**	**22.7**	**83.2**

Table 5: Figure captioning on the HCI-alt-text and SciOL.

CER↓	UBPMC	MuLMS-Img
TrOCR [36]	46.1	53.7
Sci-TrOCR	**13.5**	**14.7**

Table 6: Normalized character error rate (CER) on the UBPMC and MuLMS-Img text-recognition test subsets.

Experimental Results. Table 5 shows the captioning results of CoCa-VitB32 and SciOL-CoCa on the two test datasets. Further pre-training on SciOL leads to large improvements for both datasets. Even though CoCa-VitB32 has been trained on a shorter sequence length (77 tokens) compared to sciol (255 tokens), we observe that both models generate captions with similar sequence length of 36 tokens (HCI-alt) and 53 (SciOL) in the case of CoCa-VitB32 and 38 tokens (HCI-alt) and 48 tokens (SciOL) in the case of SciOL-CoCa. Therefore, we hypothesize that the adaptation to the domain-specific language of scientific captions is the main reason for the improvement in evaluation performance.

6.3. Optical Character Recognition Experiments

We evaluate Sci-TrOCR, our text recognition model on the test split of MuLMS-Img and on the test set of UBPMC [11]. For the latter, we use the text-recognition subsets that come with ground-truth bounding box annotations and the contained text, following the same annotation scheme as MuLMS-Img. For evaluation, we compute the character error rate (CER), normalized over the ground truth text length.

Results. Table 6 shows the results of the TrOCR model trained on SROIE [25] as well as our adapted version Sci-TrOCR, which is finetuned on MuLMS-Img and UBPMC data. Sci-TrOCR outperforms TrOCR by a large margin. We assume that the reasons for this are a better alignment with the domain-specific vocabulary as well as the dataset-specific typesetting, e.g., of mathematical formulas.

6.4. Cross-Modal Figure Retrieval Experiments

Finally, we turn to the downstream task of cross-modal figure retrieval. We test two settings: (a) In the text-to-image (t2i) setting, we use figure captions as queries, and the domain-expert written MuLMS-Img retrieval queries. (b) In the image-to-text (i2t) setting, we use images as queries with the goal to retrieve the corresponding caption. We evaluate SciOL-Coca on the SciOL test split, calculating both i2t retrieval and t2i retrieval scores. Since MuLMS-Img is composed of comparably few samples, we add 20% of the figures from the SciOL test split as negative samples and only calculate i2t scores. Following Section 5 we experiment with adding the text extracted with Sci-TrOCR from the figure (referenced as OCR) and the ground truth caption to the visual representation for the evaluation on MuLMS-Img. To evaluate retrieval performance, we calculate recall@K at ranks 1, 5 and 10, measuring the proportion of relevant images or texts that are correctly returned within the top K results.

Results. Table 7 shows figure retrieval results on the SciOL test split. Pre-training on SciOL largely improves performance compared to using non-scientific models (CoCa-ViT-B32). We also observe an improvement of 61% in text-to-image and 84% in image-to-text retrieval score at rank one compared to Biomed-CLIP. We argue that this is due to including more scientific areas compared to Biomed-CLIP.

Table 8 shows the t2i retrieval results on the MuLMS-Img queries. While we again notice the benefit of training on SciOL compared to non-scientific data (CoCa-ViT-B32), this performance increase is smaller compared to the evaluation on SciOL. Suprisingly, Biomed-CLIP performs poorer than CoCa-ViT-B32. By adding textual features, we can observe a large increase in the retrieval performance on the MuLMS-Img queries. Especially adding OCR yields gains of 13pp. in R@5 and 18pp. in R@10, nearly doubling the scores of our base SciOL-CoCa. We hypothesize that this is due to the limited implicit OCR capability of the visual encoder, especially in examples where both the query and figure explicitly mention a quantity or material. Extracting the figure text might allow more fine grained reasoning over the text with the figure. Finally, by combining pixel information, the caption and OCR, our model achieves an R@5 of 35.5%, making our system, in contrast to general-domain solutions, usable in real-world settings.

Human evaluation. Complementary to previously considering the MuLMS-Img query and figures as ground truth pairs, we conduct a human evaluation to measure the retrieval quality of all top 5 retrieved images given a search query. We task a domain expert, a graduate student of materials science with the evaluation, and calculate t2i precision and recall at rank 5. If an image is judged as relevant in this step, we add it to the relevant images for the respective query. The human annotator did not see which model performed a particular ranking. As shown in Table 8, BioMed-CLIP performs on par with CoCa-ViT-B32; SciOL-CoCa outperforms both of these models. The human evaluation also confirms the previous observation of the strong benefit of including explicit textual features from the caption and OCR in the image representations.

	text-to-image ↑			image-to-text ↑		
	R@1↑	R@5↑	R@10↑	R@1↑	R@5↑	R@10↑
CoCa-ViT-B32 [26]	1.2	2.3	3.0	0.7	1.5	1.9
BioMed-CLIP [70]	6.2	11.4	13.8	5.6	10.6	12.9
SciOL-CoCa	**10.0**	**19.1**	**23.5**	**10.3**	**19.9**	**24.6**

Table 7: Zero-shot scientific figure retrieval on the SciOL test set (using captions as queries).

	MuLMS-Img queries			human eval	
	R@1↑	R@5↑	R@10↑	R@5↑	P@5↑
BioMed-CLIP [70]	3.9	9.2	9.2	35.9	5.7
CoCa-ViT-B32 [26]	7.8	11.8	18.4	27.5	5.5
SciOL-CoCa	10.5	15.8	19.7	44.3	8.8
SciOL-CoCa text-only	10.5	13.2	14.5	-	-
SciOL-CoCa+caption	18.4	27.6	32.9	-	-
SciOL-CoCa + OCR	15.8	29.0	38.2	-	-
SciOL-CoCa+caption+OCR	**23.7**	**35.5**	**42.1**	**77.4**	**15.6**

Table 8: Zero-shot t2i retrieval on the MuLMS-Img queries. *SciOL-CoCa text-only* uses only caption and OCR texts as figure representations.

7. Conclusion

This paper has presented SciOL, the largest openly-licensed pre-training corpus for image-text models in the scientific domain. We have also introduced MuLMS-Img, a high-quality dataset from the materials science domain annotated for different text-image tasks. Our experiments have shown that pre-training multimodal models on SciOL leads to large performance gains across tasks in the scientific domain. We have also shown that with relatively little or even no fine-tuning training data for a particular domain, performance of models can be increased considerably by integrating embeddings of textual information extracted from the images. In sum, for our downstream task evaluation of cross-modal figure retrieval, integrating such features has more than doubled performance.

Potential future work could leverage the SciOL text data to build and pretrain domain specific text encoder and decoder networks. Another direction could be the explicit integration of text extracted from images into pre-training.

Acknowledgments. We thank Felix Hildebrand for valuable discussions on the materials science use cases, as well as the anonymous reviewers for their insightful comments.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, Roman Ring, Eliza Rutherford, Serkan Cabi, Tengda Han, Zhitao Gong, Sina Samangooei, Marianne Monteiro, Jacob Menick, Sebastian Borgeaud, Andrew Brock, Aida Nematzadeh, Sahand Sharifzadeh, Mikolaj Binkowski, Ricardo Barreira, Oriol Vinyals, Andrew Zisserman, and Karen Simonyan. Flamingo: a visual language model for few-shot learning. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022.

[2] Abhijit Balaji, Thuvaarakkesh Ramanathan, and Venkateshwarlu Sonathi. Chart-text: A fully automated chart image descriptor, 2018.

[3] Hangbo Bao, Wenhui Wang, Li Dong, Qiang Liu, Owais Khan Mohammed, Kriti Aggarwal, Subhojit Som, and Furu Wei. Vlmo: Unified vision-language pre-training with mixture-of-modality-experts. In *2021 Neural Information Processing Systems*, November 2021.

[4] Steven Bird, Ewan Klein, and Edward Loper. *Natural Language Processing with Python: Analyzing Text with the Natural Language Toolkit*. O'Reilly Media, Inc., 2009.

[5] Lutz Bornmann, Robin Haunschild, and Rüdiger Mutz. Growth rates of modern science: a latent piecewise growth curve approach to model publication numbers from established and new literature databases. *Palgrave Communications*, 8(1):1–15, December 2021.

[6] Ritwick Chaudhry, Sumit Shekhar, Utkarsh Gupta, Pranav Maneriker, Prann Bansal, and Ajay Joshi. Leaf-qa: Locate, encode & attend for figure question answering. *2020 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 3501–3510, 2019.

[7] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. Uniter: Universal image-text representation learning. In *European Conference on Computer Vision*, 2019.

[8] Sanjana Shivani Chintalapati, Jonathan Bragg, and Lucy Lu Wang. A dataset of alt texts from HCI publications: Analyses and uses towards producing more descriptive alt texts of data visualizations in scientific papers. In *Proceedings of the 24th International ACM SIGACCESS Conference on Computers and Accessibility*. Association for Computing Machinery, Oct. 2022.

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

[10] Ekin Dogus Cubuk, Barret Zoph, Jonathon Shlens, and Quoc V. Le. Randaugment: Practical automated data augmentation with a reduced search space. *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 3008–3017, 2019.

[11] Kenny Davila, Fei Xu, Saleem Ahmed, David A. Mendoza, Srirangaraj Setlur, and Venu Govindaraju. Icpr 2022: Challenge on harvesting raw tables from infographics (chart-infographics). In *2022 26th International Conference on Pattern Recognition (ICPR)*, pages 4995–5001, 2022.

[12] Karan Desai and Justin Johnson. Virtex: Learning visual representations from textual annotations. *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11157–11168, 2020.

[13] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics, 2019.

[14] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *ArXiv*, abs/2010.11929, 2020.

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

[16] Sedigheh Eslami, Christoph Meinel, and Gerard de Melo. Pubmedclip: How much does CLIP benefit visual question answering in the medical domain? In Andreas Vlachos and Isabelle Augenstein, editors, *Findings of the Association for Computational Linguistics: EACL 2023, Dubrovnik, Croatia, May 2-6, 2023*, pages 1151–1163. Association for Computational Linguistics, 2023.

[17] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Proceedings of the Second International Conference on Knowledge Discovery and Data Mining*, KDD'96, page 226–231. AAAI Press, 1996.

[18] Annemarie Friedrich, Heike Adel, Federico Tomazic, Johannes Hingerl, Renou Benteau, Anika Maruscyk, and Lukas Lange. The sofc-exp corpus and neural approaches to information extraction in the materials science domain. In *Annual Meeting of the Association for Computational Linguistics*, 2020.

[19] José Manuél Gómez-Pérez and Raúl Ortega. Look, read and enrich - learning from scientific figures and their captions. *Proceedings of the 10th International Conference on Knowledge Capture*, 2019.

[20] Yu Gu, Robert Tinn, Hao Cheng, Michael R. Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3:1 – 23, 2020.

[21] Kaiming He, Haoqi Fan, Yuxin Wu, Saining Xie, and Ross B. Girshick. Momentum contrast for unsupervised visual representation learning. *2020 IEEE/CVF Conference*

on Computer Vision and Pattern Recognition (CVPR), pages 9726–9735, 2019.

[22] Nejc Hodnik, Luigi Romano, Primož Jovanovič, Francisco Ruiz-Zepeda, Marjan Bele, Filippo Fabbri, Luana Persano, Andrea Camposeo, and Dario Pisignano. Assembly of pt nanoparticles on graphitized carbon nanofibers as hierarchically structured electrodes. *ACS Applied Nano Materials*, 3(10):9880–9888, 2020. PMID: 33134881.

[23] Ting-Yao Hsu, C. Lee Giles, and Ting-Hao 'Kenneth' Huang. Scicap: Generating captions for scientific figures. In *Conference on Empirical Methods in Natural Language Processing*, 2021.

[24] Zhi Huang, Federico Bianchi, Mert Yuksekgonul, Thomas Montine, and James Zou. Leveraging medical twitter to build a visual–language foundation model for pathology ai. *bioRxiv*, 2023.

[25] Zheng Huang, Kai Chen, Jianhua He, Xiang Bai, Dimosthenis Karatzas, Shijian Lu, and C. V. Jawahar. Icdar2019 competition on scanned receipt ocr and information extraction. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1516–1520, 2019.

[26] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, July 2021. If you use this software, please cite it as below.

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

[28] Ron Daniel Jr., Sujit Pal, and Andrew Epstein. OA-STM-Corpus. `https://github.com/elsevierlabs/OA-STM-Corpus`, 2017.

[29] Kushal Kafle, Scott D. Cohen, Brian L. Price, and Christopher Kanan. Dvqa: Understanding data visualizations via question answering. *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 5648–5656, 2018.

[30] Wonjae Kim, Bokyung Son, and Ildoo Kim. Vilt: Vision-and-language transformer without convolution or region supervision. In *International Conference on Machine Learning*, 2021.

[31] Harold W. Kuhn. *The Hungarian Method for the Assignment Problem*, pages 29–47. Springer Berlin Heidelberg, Berlin, Heidelberg, 2010.

[32] Jason J Lau, Soumya Gayen, Dina Demner, and Asma Ben Abacha. Visual question answering in radiology (vqa-rad), Feb 2019.

[33] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *ArXiv*, abs/2301.12597, 2023.

[34] Junnan Li, Dongxu Li, Caiming Xiong, and Steven C. H. Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, 2022.

[35] Junnan Li, Ramprasaath R. Selvaraju, Akhilesh Deepak Gotmare, Shafiq Joty, Caiming Xiong, and Steven Hoi. Align before fuse: Vision and language representation learning with momentum distillation. In *NeurIPS*, 2021.

[36] Minghao Li, Tengchao Lv, Lei Cui, Yijuan Lu, Dinei Florencio, Cha Zhang, Zhoujun Li, and Furu Wei. Trocr: Transformer-based optical character recognition with pre-trained models, 2021.

[37] Yanghao Li, Haoqi Fan, Ronghang Hu, Christoph Feichtenhofer, and Kaiming He. Scaling language-image pre-training via masking. In *CVPR*, 2023.

[38] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics.

[39] Tsung-Yi Lin, Piotr Dollar, Ross Girshick, Kaiming He, Bharath Hariharan, and Serge Belongie. Feature pyramid networks for object detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, July 2017.

[40] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft coco: Common objects in context. In David Fleet, Tomas Pajdla, Bernt Schiele, and Tinne Tuytelaars, editors, *Computer Vision – ECCV 2014*, pages 740–755, Cham, 2014. Springer International Publishing.

[41] Weixiong Lin, Ziheng Zhao, Xiaoman Zhang, Chaoyi Wu, Ya Zhang, Yanfeng Wang, and Weidi Xie. Pmc-clip: Contrastive language-image pre-training using biomedical documents. *ArXiv*, abs/2303.07240, 2023.

[42] Pan Lu, Swaroop Mishra, Tony Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *The 36th Conference on Neural Information Processing Systems (NeurIPS)*, 2022.

[43] Jason Obeid and Enamul Hoque. Chart-to-text: Generating natural language descriptions for charts by adapting the transformer model. In *Proceedings of the 13th International Conference of Natural Language Generation*, pages 138–147, Dublin, Ireland, Dec. 2020. Association for Computational Linguistics.

[44] Department of Health and Human Services. Index of /pub-/pmc/oa_bulk/oa_noncomm. `https://ftp.ncbi.nlm.nih.gov/pub/pmc/oa_bulk/oa_noncomm/`, 2023.

[45] Directory of Open Access Journals. Directory of Open Access Journals Open Archives Initiative Endpoint. `https://doaj.org/oai`, 2023.

[46] Alin Orfanidi, Philipp J. Rheinländer, Nicole Schulte, and Hubert A. Gasteiger. Ink solvent dependence of the ionomer distribution in the catalyst layer of a pemfc. *Journal of The Electrochemical Society*, 165(14):F1254, nov 2018.

[47] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In Marina Meila

and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 2021.

[48] Shaoqing Ren, Kaiming He, Ross Girshick, and Jian Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In C. Cortes, N. Lawrence, D. Lee, M. Sugiyama, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 28. Curran Associates, Inc., 2015.

[49] Robin Rombach, A. Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10674–10685, 2021.

[50] Johannes Rückert, Asma Ben Abacha, Alba García Seco de Herrera, Louise Bloch, Raphael Brüngel, Ahmad Idrissi-Yaghir, Henning Schäfer, Henning Müller, and C. Friedrich. Overview of imageclefmedical 2022 - caption prediction and concept detection. In *Conference and Labs of the Evaluation Forum*, 2022.

[51] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Raphael Gontijo-Lopes, Burcu Karagol Ayan, Tim Salimans, Jonathan Ho, David J. Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022.

[52] Timo Pierre Schrader, Matteo Finco, Stefan Grünewald, Felix Hildebrand, and Annemarie Friedrich. Mulms: A multilayer annotated text corpus for information extraction in the materials science domain, 2023.

[53] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade W Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa R Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. LAION-5b: An open large-scale dataset for training next generation image-text models. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2022.

[54] Zejiang Shen, Kyle Lo, Lucy Lu Wang, Bailey Kuehl, Daniel S Weld, and Doug Downey. Vila: Improving structured content extraction from scientific pdfs using visual layout groups. *Transactions of the Association for Computational Linguistics*, 10:376–392, 2022.

[55] Zejiang Shen, Ruochen Zhang, Melissa Dell, Benjamin Charles Germain Lee, Jacob Carlson, and Weining Li. Layoutparser: A unified toolkit for deep learning based document image analysis. *arXiv preprint arXiv:2103.15348*, 2021.

[56] Jeremy Singer-Vine. pdfplumber. `https://github.com/jsvine/pdfplumber`, 2023.

[57] Hrituraj Singh and Sumit Shekhar. STL-CQA: Structure-based transformers with localization and encoding for chart question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3275–3284, Online, Nov. 2020. Association for Computational Linguistics.

[58] Yufei Song, Wei Wang, Lei Ge, Xiaomin Xu, Zhenbao Zhang, Paulo Sérgio Barros Julião, Wei Zhou, and Zongping Shao. Rational design of a water-storable hierarchical architecture decorated with amorphous barium oxide and nickel nanoparticles as a solid oxide fuel cell anode with excellent sulfur tolerance. *Advanced Science*, 4(11):1700337, 2017.

[59] Hao Tan and Mohit Bansal. LXMERT: Learning cross-modality encoder representations from transformers. 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 5100–5111, Hong Kong, China, Nov. 2019. Association for Computational Linguistics.

[60] Tzutalin. LabelImg. Git code. `https://github.com/tzutalin/labelImg`, 2015.

[61] Cornell University, Joe Tricot, devrishi, Brian Maltzan, and Shamsi Brinn. arXiv Dataset. `https://www.kaggle.com/datasets/Cornell-University/arxiv`, 2023.

[62] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008, 2017.

[63] Bastiaan S. Veeling, Jasper Linmans, Jim Winkens, Taco Cohen, and Max Welling. Rotation equivariant cnns for digital pathology, 2018.

[64] Peng Wang, An Yang, Rui Men, Junyang Lin, Shuai Bai, Zhikang Li, Jianxin Ma, Chang Zhou, Jingren Zhou, and Hongxia Yang. Unifying architectures, tasks, and modalities through a simple sequence-to-sequence learning framework. In *International Conference on Machine Learning*, 2022.

[65] Zifeng Wang, Zhenbang Wu, Dinesh Agarwal, and Jimeng Sun. Medclip: Contrastive learning from unpaired medical images and text. *ArXiv*, abs/2210.10163, 2022.

[66] Zirui Wang, Jiahui Yu, Adams Wei Yu, Zihang Dai, Yulia Tsvetkov, and Yuan Cao. SimVLM: Simple visual language model pretraining with weak supervision. In *International Conference on Learning Representations*, 2022.

[67] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *Transactions on Machine Learning Research*, 2022.

[68] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, Benton C. Hutchinson, Wei Han, Zarana Parekh, Xin Li, Han Zhang, Jason Baldridge, and Yonghui Wu. Scaling autoregressive models for content-rich text-to-image generation. *Trans. Mach. Learn. Res.*, 2022, 2022.

[69] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao.

Vinvl: Revisiting visual representations in vision-language models. In *2021 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 5575–5584, 2021.

[70] Sheng Zhang, Yanbo Xu, Naoto Usuyama, Jaspreet Bagga, Robert Tinn, Sam Preston, Rajesh Rao, Mu Wei, Naveen Valluri, Cliff Wong, Matthew P. Lungren, Tristan Naumann, and Hoifung Poon. Large-scale domain-specific pretraining for biomedical vision-language processing. *CoRR*, abs/2303.00915, 2023.

[71] Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*, 2020.

[72] Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D. Manning, and C. Langlotz. Contrastive learning of medical visual representations from paired images and text. *ArXiv*, abs/2010.00747, 2020.

[73] Xu Zhong, Jianbin Tang, and Antonio Jimeno Yepes. Publaynet: largest dataset ever for document layout analysis. In *2019 International Conference on Document Analysis and Recognition (ICDAR)*, pages 1015–1022. IEEE, Sep. 2019.

[74] Luowei Zhou, Hamid Palangi, Lei Zhang, Houdong Hu, Jason J. Corso, and Jianfeng Gao. Unified vision-language pre-training for image captioning and vqa. *ArXiv*, abs/1909.11059, 2019.