mOSCAR: A LARGE-SCALE MULTILINGUAL AND MULTIMODAL DOCUMENT-LEVEL CORPUS

Anonymous authors
Paper under double-blind review

ABSTRACT

Multimodal Large Language Models (mLLMs) are trained on a large amount of text-image data. While most mLLMs are trained on caption-like data only, Alayrac et al. (2022) showed that additionally training them on interleaved sequences of text and images can lead to the emergence of in-context learning capabilities. However, the dataset they used, M3W, is not public and is only in English. There have been attempts to reproduce their results but the released datasets are English-only. In contrast, current multilingual and multimodal datasets are either composed of caption-like only or medium-scale or fully private data. This limits mLLM research for the 7,000 other languages spoken in the world. We therefore introduce mOSCAR, to the best of our knowledge the first large-scale multilingual and multimodal document corpus crawled from the web. It covers 163 languages, 303M documents, 200B tokens and 1.15B images. We carefully conduct a set of filtering and evaluation steps to make sure mOSCAR is sufficiently safe, diverse and of good quality. We additionally train two types of multilingual model to prove the benefits of mOSCAR: (1) a model trained on a subset of mOSCAR and captioning data and (2) a model trained on captioning data only. The model additionally trained on mOSCAR shows a strong boost in few-shot learning performance across various multilingual image-text tasks and benchmarks, confirming previous findings for English-only mLLMs. The dataset will be made publicly accessible.

1 INTRODUCTION

Multimodal large language models (mLLMs) are trained on large amounts of text-image data (Radford et al., 2021; Yu et al., 2022; Li et al., 2023; Wang et al., 2023; OpenAI, 2023; Gemini Team et al., 2023; Chameleon Team, 2024). The main paradigm until recently was to train a model from a large collection of web-crawled images and their captions (Li et al., 2021; Wang et al., 2022; Chen et al., 2023b). Models such as Flamingo (Alayrac et al., 2022) challenged this paradigm by being additionally trained on interleaved sequences of text and images from web documents, showing state-of-the-art results on various tasks and in-context learning capabilities that are not present in models trained on caption-like data only. Additionally, McKinzie et al. (2024) recently proved that including interleaved text-image data during training was necessary to get good few-shot learning performance. However, the datasets used to train mLLMs are either private (Alayrac et al., 2022), monolingual or multilingual but only medium-scale (Srinivasan et al., 2021). Some attempts have been made to reproduce these datasets (Zhu et al., 2023; Laurençon et al., 2023) but the resulting datasets are only available in English.

Few image-text datasets are multilingual and most of them are obtained by translating English caption-like datasets, such as multilingual Conceptual Captions (Sharma et al., 2018), into multiple languages using neural machine translation (NMT) systems (Surís et al., 2022; Maaz et al., 2024). This presents some drawbacks such as some languages still being poorly translated by current state-of-the-art NMT models (Liu et al., 2020; Costa-jussà et al., 2022) and some cultural subtleties inherent in each language not being fully conveyed. Some efforts have been conducted to collect large-scale multilingual image captioning datasets, such as LAION-5B (Schuhmann et al., 2022), but they are limited to caption data too, are relatively noisy and more importantly contain a non-negligible share of "not safe for work" (NSFW) content such as pædopornographic images (Schuhmann et al., 2022).

Le plus charmant petit square de Paris.

Le square René Viviani a été ouvert au public en 1928. Il occupait l'emplacement de l'annexe de l'Hôtel-Dieu (il est situé en face de Notre Dame) et d'une rangée de maisons longeant la rue de la Bûcherie. Le conseiller municipal Léon Riotor et la Commission du Vieux Paris ont particulièrement oeuvré à sa création dans le but de préserver à la fois le dégagement de l'église Saint-Julien-le-Pauvre et la perspective de Notre-Dame. Quelques éléments architecturaux parsèment le parc, et contribuent à son charme : pinacles, chapiteaux, balustrades, provenant sans doute de la cathédrale Notre-Dame. [...]



Pinacles et chapiteaux ornent le parc : probablement des vestiges de Notre-Dame. La fine poussière derrière la balustrade de pierre provient du fort vent qui faisait s'envoler le sable et la terre. [...]

Figure 1: Example of a French document from mOSCAR.

This motivated us to collect and release the first large-scale multilingual and multimodal document dataset derived from Common Crawl.[1] Our dataset, multimodal OSCAR (mOSCAR), follows the OSCAR initiative (Ortiz Suárez et al., 2019; Abadji et al., 2021; 2022) and covers 303M documents in 163 languages, 200B tokens and 1.15B images. Figure 1 shows an example of a document, more can be found in the Supplementary Material. We carry out extensive filtering to increase its safety and quality. To prove mOSCAR's utility, we train a multilingual OpenFlamingo (Awadalla et al., 2023) from a Gemma-2B language model (Gemma Team et al., 2024) on a subset of mOSCAR and captioning data from LAION-400M (Schuhmann et al., 2021), recaptioned with BLIP (Li et al., 2022), filtered with CLIP (Radford et al., 2021) and translated with NLLB (Costa-jussà et al., 2022). We compare against a similar model trained on captioning data only and show we obtain a strong boost in few-shot learning, confirming previous findings for English (Alayrac et al., 2022; McKinzie et al., 2024; Laurençon et al., 2024). The dataset and models will be made publicly available.

2 Related Work

Large-scale web-based datasets Numerous datasets have been created by filtering web-crawled data. These include large-scale text-only datasets (Ortiz Suárez et al., 2019; Raffel et al., 2020; Wenzek et al., 2020; Gao et al., 2020; Abadji et al., 2021; Xue et al., 2021; Laurençon et al., 2022; Abadji et al., 2022; Penedo et al., 2023) and multimodal ones (Sharma et al., 2018; Changpinyo et al., 2021; Jia et al., 2021; Schuhmann et al., 2021; 2022; Byeon et al., 2022; Laurençon et al., 2023; Zhu et al., 2023; Gadre et al., 2024). Even if these datasets are not as high quality as smaller and/or hand-crafted ones, they are now the standard to pretrain foundation models, as it has been shown that training bigger models on more data leads to better downstream performances (Brown et al., 2020; Hoffmann et al., 2022; Touvron et al., 2023a;b).

English image-text datasets The first open-source image-text datasets were manually created, small-scale and English-only (Ordonez et al., 2011; Lin et al., 2014; Plummer et al., 2015; Krishna et al., 2017). Scaling up these datasets was an appealing solution to overcome limitations of previous image-text models; a few works (Sharma et al., 2018; Changpinyo et al., 2021) proposed to collect millions of image-text pairs from the web before filtering them with well-designed steps. Relaxing the filtering steps enabled the collection of more data and led to large-scale datasets to train image-text foundation models (Radford et al., 2021; Li et al., 2021; Schuhmann et al., 2021; 2022; Byeon et al., 2022). However, these datasets generally contain caption-like image-text pairs only, and it is therefore difficult to observe in-context learning abilities similarly to text-only language models trained on raw documents (Raffel et al., 2020). Alayrac et al. (2022) overcome this issue by training their model directly on documents with interleaved image-text data. While their results are promising, their M3W dataset is English-only and private. Recently, open-source efforts (Zhu et al., 2023; Laurençon et al., 2023) have been made to release a similar dataset but they are still monolingual.

[1] `https://commoncrawl.org/`. The Common Crawl Foundation is a non-profit organization that crawls the web on a monthly basis.

Multilingual image-text datasets Only a few image-text datasets are available in multiple languages. One of the first focused on collecting Google images from short queries based on word frequencies from Wikipedia pages in 98 languages (Hewitt et al., 2018). Later, Srinivasan et al. (2021) proposed the WIT dataset, an image-text dataset composed of Wikipedia pages. Although of high quality, it is only medium-scale even for high-resource languages and there are fewer than 50k unique images for most languages. Another approach lies in bootstrapping multilingual and multimodal data from a model trained with English-only data (Mohammed et al., 2023). While effective for captioning, it is computationally expensive to implement in practice. Other multilingual image-text datasets exist but focus on captions only and are highly domain-specific (Kosar et al., 2022; Leong et al., 2022).

3 Dataset Creation Pipeline

3.1 Data collection

We collect mOSCAR from the Web ARchive Content (WARC) files of three 2023 Common Crawl dumps, processing them using the FastWARC library (Bevendorff et al., 2021). We remove documents smaller than 500 bytes (50% of the documents), as we find they are usually too small to be considered documents and tend to contain noisy text. We then navigate through the entire Document Object Model (DOM) tree with a depth first search algorithm and ChatNoir library (Bevendorff et al., 2018) to extract nodes of interests corresponding to specific HTML tags.

Following previous work, we extract text from the tags that usually contain the main content of web pages (we refer to them as DOM text nodes), i.e. `<p>`, `<h*>`, `<title>`, `<description>`, `<ul>`, `<ol>`, `<aside>`, `<dl>`, `<dd>`, `<dt>`. Similarly to (Laurençon et al., 2023), we choose to remove `<table>` content as most often it is irrelevant and difficult to render. We extract all `<img>` tags (we refer to them as DOM image nodes). We then remove documents with fewer than 3 text nodes (as they do not contain enough text) and more than 30 image nodes (as we found them to be too noisy).

3.2 Language identification

We identify the language of each document using the state-of-the-art open-LID language detector (Burchell et al., 2023), covering 201 languages. We apply open-LID to each DOM text node and keep the three most probable languages with their respective probabilities. The language of the document is then determined by summing over the probabilities of each language detected for each text segment, weighted by the number of characters in the segment[2] and taking the language with the highest score.

3.3 Text-only filtering

We apply a series of filtering steps to the text content of each document independently of the images, with the aim of discarding poor quality documents and cleaning text as best as possible. We first filter at the text-node level and then at the whole document level, before running near-deduplication to keep unique text nodes within a document and unique documents in the dataset.

Text node filtering We use a set of heuristics (see Supplementary Material) to extract as much human-generated content as possible while discarding noisy text related to ads and website functions (e.g. “Instagram”, “Facebook”). We then keep DOM text nodes with content over 10 bytes. This step, designed to improve the quality of extracted text, removes on average 55% of text nodes.

Document filtering We mostly filter “not safe for work” (NSFW) content at the document level. We use an English regular expression to detect adult content, similar to the one used by the Université Toulouse 1 Capitole[3] and remove the entire document if there is a match with any of the DOM text nodes’ contents, removing on average 0.5% of documents (mostly English ones). We acknowledge that there is a high probability that this also discards safe content, e.g. we could remove content from

[2]This is to avoid mis-assigning the language due to the presence of many short, non-informative DOM text nodes in the same language (e.g. “Cookies”, “Subscribe”, “Newsletter” etc.) and because language identification is generally less reliable for short segments.

[3]`https://dsi.ut-capitole.fr/blacklists/index_en.php`

certain communities who use some explicit words in a non-sexual way (Sap et al., 2019). However, we explicitly favour recall over precision to minimise the risk of unsafe content. We additionally remove documents containing fewer than five DOM text nodes and fewer than 300 characters after the previous filtering steps, removing 70.6% of documents.

Deduplication We conduct several types of per-language deduplication at different levels, as this has been shown to improve training efficiency (Abbas et al., 2023). First, we keep unique documents only by removing exact duplicates at the document level. We also remove exact duplicates of text nodes within the same document (4% of text nodes) and near-duplicate text nodes (1% of text nodes) by computing the Levenshtein ratio (Levenshtein, 1966) between all text nodes within the same document and applying a threshold of 0.95. If near-duplicates are found, we keep the first one in the document. Finally, we conduct per language near-deduplication at the document level with MinHashLSH (Broder, 1997; Gionis et al., 1999) following Smith et al. (2022), removing on average 19% of documents:[4] we turn documents into hashing vectors, compute min hashes from these vectors and perform Locality Sensitive Hashing to remove duplicates[5] (see Supplementary Material for more details).

Toxicity filtering Toxic content targeting individuals or groups of people is widespread on the internet and can therefore be found in large-scale web-crawled datasets like mOSCAR without appropriate filtering steps. To alleviate this issue, we apply the same method used by Costa-jussà et al. (2022) and remove documents from mOSCAR based on the presence of a list of "toxic" words for each language[6]. As some words in the list can also be used in a non-toxic way based on the context (e.g.: 'breast' in English), we tag the document as toxic and remove it from mOSCAR if it contains at least two distinct words in the list. This filtering step removes 0.95% of the documents for very high-resource languages (>5M documents), 2.13% for high-resource languages (<5M, >500K), 0.47% for mid-resource languages (<500K, >50K) and 0.64% for low-resource languages (< 50K). When manually analysing 100 random documents removed by this filtering step in each of the 2 (high-resource) languages we are native speakers of (English and French), we found 58 documents with toxic content.

Personal Identifiable Information Personal Identifiable Information (PII) can be found in large-scale web-crawled datasets, we therefore conducted an additional filtering step to replace all detected PII by place holder strings using regular expressions (see Supplementary Material for the list of regular expressions we used). Concretely, we replaced all detected email addresses, phone numbers, credit card numbers, IP addresses and passport numbers.

3.4 IMAGE-ONLY FILTERING

We downloaded images from the URLs in DOM image nodes using a modified version of the img2dataset toolkit (Beaumont, 2021) that includes an antivirus scan and follows `robots.txt` instructions to respect the Robots Exclusion Protocol. We then apply a series of filtering steps, first removing images based on heuristics, and then applying multiple NSFW detection models to remove undesirable content. Finally, we conduct a set of deduplication steps.

Rule-based filters Similarly to previous works (Schuhmann et al., 2021) and to avoid extracting low-resolution images and favicons, we keep images with a minimum height and width of 150 pixels. We restrict the aspect ratio to be between 3 and 1/3 (to remove banners), we remove images if their URLs contain the words "logo", "banner", "button", "widget", "icon" or "plugin" or if the image name from the URL matches "twitter", "facebook" or "rss" (to remove logos). This step removes 13.6% of the URLs. At this stage, we downloaded 2.5B images with an average success rate of 55%.

NSFW detection We use multiple NSFW automatic models to remove as much unsafe content as possible. We first combine two NSFW detectors: nsfw-detector (Laborde), a 5-class classifier with a

[4]With some disparity among languages as we found more duplicates for low- than high-resource languages.

[5]We performed this using the `datasketch` python library.

[6]The list of these words for each language can be found here: `https://github.com/facebookresearch/flores/tree/main/toxicity`

MobileNet (Howard et al., 2017) backbone fine-tuned on 60GB of annotated data and NudeNet,[7] an object detector trained to detect different types of nudity in images. We combined the two models as we found the first to be gender-biased while the second gives a large number of false positives for non-human images. Concretely, we consider an image an NSFW candidate if the sum of the probabilities for the classes 'porn' and 'hentai' is superior to 0.8 using nsfw-detector. We then tag the image as NSFW if one of the sensitive 'exposed' classes of NudeNet gets a probability superior to 0.5. If a document contains an image with an NSFW tag, we remove the entire document from the dataset, which removes 0.5% of images. We manually inspecting 1,000 images of the remaining data and found no NSFW content. We manually inspected 1,000 images of the removed content and found 63.4% of NSFW images.

CSAM content Child Sexual Abuse Material (CSAM) is widespread on the internet and is therefore likely to be found in such a large-scale dataset crawled from the web. Removing CSAM is challenging as there is no training data nor open-source detection models available as these could be used in a harmful way. We rely on Thorn's CSAM classifier[8], a proprietary classifier trained to detect CSAM content in images. We tag the image as CSAM if the probability of the class CSAM is superior to 0.4 to favour recall over precision. As mentioned above, if a document contains an image with a CSAM tag, we remove it from the dataset. This step removes 0.07% of the images.

Deduplication To avoid memorisation issues often seen in models trained on datasets with many duplicated images (Somepalli et al., 2023; Carlini et al., 2023; Webster et al., 2023; Somepalli et al., 2024), we perform deduplication at the image level. We first remove duplicate images within the same document by URL matching (removing 8.7% of URLs). We then compute a perceptual hash (pHash) for each image using the imagehash library[9] and remove images with the same pHash within the same document, keeping only the first occurrence. We also limit the number of times an image can appear in the dataset per-language to 10 using both URL matching and perceptual hashing (this removes 2.5% of images). We do this per-language and not across languages as having the same images in documents from different languages could encourage cross-lingual transfer.

Personal Identification Information To protect PII in images, we use a lightweight face detector[10] and apply a threshold of 0.99 to detect faces in the images. We apply such a high threshold as we found the model to be biased towards detecting faces with high probability in images without any human. For each image in mOSCAR, we distribute the bounding boxes of the detected faces so that users can blur them when downloading the images. More details are provided in the Supplementary Material.

3.5 Data decontamination

LLMs and mLLMs are trained on web-crawled data that can contain the benchmarks they are tested on (Dodge et al., 2021). As they are good at memorizing training data (Carlini et al., 2023), this data contamination is problematic. We therefore discard all images with the same perceptual hash as any of the images from the evaluation benchmarks (and their training sets) we use (see Section 5.1). This step removes on average 126,016 images for high-resource languages (up to 300K images for English), 6,862 images for mid-resource languages and 45 images for low-resource languages.

3.6 Text-image joint filtering

Our aim is to obtain truly multimodal documents where all images are related to at least one of the text nodes in some way[11] and vice versa. We choose to apply joint text-image filtering to discard images and/or text nodes that are irrelevant to the rest of the document (e.g. the case of ads and website functionalities). To do this, we use NLLB-SIGLIP[12] (Visheratin, 2023), a multilingual version

[7] `https://github.com/vladmandic/nudenet`

[8] `https://safer.io/`

[9] `https://github.com/JohannesBuchner/imagehash`

[10] `https://github.com/Linzaer/Ultra-Light-Fast-Generic-Face-Detector-1MB`

[11] We do not limit ourselves to caption-like relation and instead allow all types of text-image relation.

[12] `siglip-base-patch16-224` as vision encoder and `nllb-distilled-600M` as text encoder.

of SIGLIP (Zhai et al., 2023) trained with the encoder of NLLB (Costa-jussà et al., 2022), which covers all mOSCAR languages.[13] We compute cosine similarity scores between all images and all paragraphs[14] within a same document. To remove irrelevant text nodes or images in a document, we mimic a text-image retrieval task, which means we avoid using arbitrary cosine similarity thresholds for each language and can reduce length biases and those in favour of caption-like paragraphs. For each candidate pair we randomly sample 63 negative images and 63 negative similar-length paragraphs from the same language but other documents. We tag the text node (resp. image) as valid if the cosine similarity of the pair is among the top 8 of the text-to-image (resp. image-to-text) similarity scores computed with the candidate text node (resp. image) and all the negative images (resp. text nodes). This means that we tag the text node (resp. image) as valid if it has a significantly higher score than a score computed with a random image (resp. text) for at least one of the images (resp. text node) in the document. We then discard text nodes and images not tagged as valid (on average 35% of the DOM text nodes and 10% of the images within a document).

After this filtering step, we apply additional text-only filters to keep documents superior to 100 bytes. We also reapply the open-lid language detector (Burchell et al., 2023) as described in Section 3.2 as we found the last filtering step to change the major language of some documents.

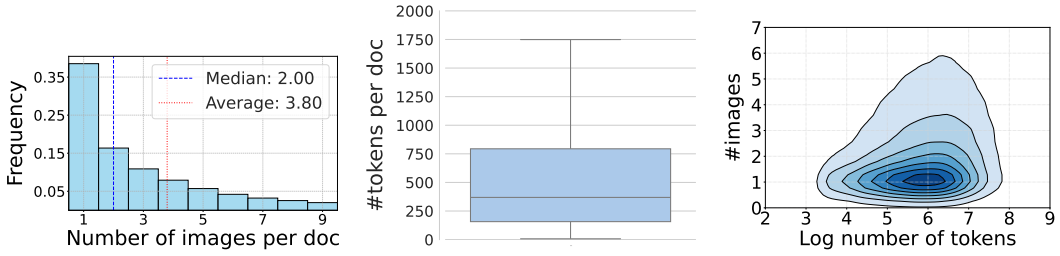
4 Multimodal Open Super-large Crawled Aggregated coRpus (mOSCAR)

mOSCAR is extracted from three Common Crawl dumps from 2023. Due to computational constraints and in order to extract a maximum number of documents for low-resource languages, we extracted all languages from the first dump only. We removed the 6 most high-resource languages from the second dump and only extracted the languages with fewer than 1M documents for the last dump. Table 1 shows a distribution of the total number of languages and their number of documents.

To avoid data poisoning (Carlini et al., 2024), we release a hash (sha512) with each mOSCAR image.

#documents	10M	5M	1M	500K	200K	50K	10K	5K	1K
#languages	10	15	36	46	56	75	119	133	154

Table 1: Number of languages with at least N documents



(a) Number of images (b) Number of tokens (c) Number of images and tokens

Figure 2: Distributions of numbers of tokens and images per document

mOSCAR is composed of 303M documents (200B tokens, 1.15B images) from 163 languages. Figure 2 shows the distribution of images and tokens per document and their joint distribution. As shown in Figure 2a, the mean and median number of images per document is 2 and 3.80.

4.1 Quality vs Diversity

While improving overall data quality, the filtering steps we applied (see Section 3) necessarily have a negative impact on diversity. We therefore study the trade-off between quality and diversity and compare against previously published, well-used datasets.

[13]We use the open-clip (Ilharco et al., 2021) model version and the transformers (Wolf et al., 2020) library.

[14]We refer to paragraph as the text content in a DOM text node.

4.1.1 TEXT CONTENT

Diversity By contruction, mOSCAR is diverse in terms of number of languages, so we focus on the diversity of mOSCAR's English documents and compare against mmc4 (Zhu et al., 2023), OBELICS (Laurençon et al., 2023) and the English subset of WIT (Srinivasan et al., 2021). We compute the Vendi score (Friedman and Dieng, 2023) on a set of SimCSE embeddings (Gao et al., 2021) with a RoBERTa encoder (Liu et al., 2019) to evaluate the content diversity. Since embedding-based diversity metrics target content diversity well but are less relevant for lexical diversity (Tevet and Berant, 2021), we measure lexical diversity via the distinct n-gram ratio (Li et al., 2016).

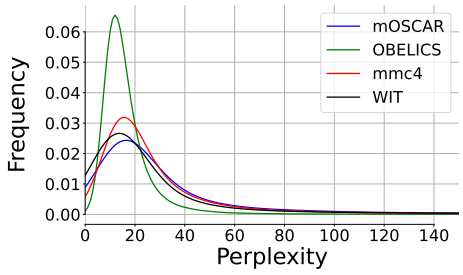


Figure 3: Perplexity of 100K random documents from different datasets.

	Vendi score	Dist. n-gram ratio
mOSCAR	<u>69.05</u> ($\pm$ 0.14)	0.472 ($\pm$ 0.002)
mmc4	67.93 ($\pm$ 0.12)	<u>0.494</u> ($\pm$ 0.002)
OBELICS	58.49 ($\pm$ 0.09)	0.488 ($\pm$ 0.001)
WIT	**73.30** ($\pm$ 0.09)	**0.530** ($\pm$ 0.001)

Table 2: Average text diversity scores ($\pm$ standard error) of text documents.

Comparison with other datasets For content diversity, we randomly sample 30M documents for mOSCAR, mmc4 and OBELICS and 3M documents for WIT and represent the documents by their SimCSE embedding. We compute the Vendi Score with cosine similarity on a randomly sampled subset of 65,536 documents. Table 2 shows that mOSCAR English content is more diverse than mmc4 and OBELICS but less diverse than WIT. For lexical diversity, we randomly sample 3M documents for mOSCAR, mmc4, OBELICS and WIT and compute the distinct n-gram ratio on a subset of 8,192 documents for n from 1 to 4. Table 2 shows that mOSCAR is slightly less lexically diverse than OBELICS and mmc4, while WIT is by far the most diverse.

Quality To evaluate document quality, we focus on English documents and compute their perplexity using Gemma-2B (Gemma Team et al., 2024). Figure 3 shows the kernel density estimation of the distribution of the perplexity of 100K randomly sampled documents from different datasets: mOSCAR is comparable to mmc4 and WIT, while OBELICS appears to be the of the highest quality. mOSCAR is therefore comparable to other interleaved image-text dataset in terms of quality and diversity of its English subset. It is however more diverse than English-only datasets by its multilingual construction and more than 10 times larger than existing multilingual interleaved image-text datasets such as WIT.

4.1.2 IMAGE DIVERSITY

mOSCAR	LAION-400M	WIT
<u>55.74</u> ($\pm$ 0.16)	**67.59** ($\pm$ 0.16)	36.14 ($\pm$ 0.08)

(a) Comparison of different datasets.

English	All
52.36 ($\pm$ 0.18)	**54.78** ($\pm$ 2.29)

(b) mOSCAR (English vs. any language).

Table 3: Average Vendi score ($\pm$ standard error) of images.

Comparison with other datasets We compute the Vendi Score on random samples of images for different datasets, comparing the images from English mOSCAR documents with those from Conceptual Captions (Changpinyo et al., 2021), LAION-400M (Schuhmann et al., 2021) and WIT (Srinivasan et al., 2021). We represent each image by its SigLIP[15] (Zhai et al., 2023) embedding and compute the Vendi score on batches of size 65,536 and a total of 1M images for each dataset. In

[15]We use `siglip-base-patch16-224`.

Table 3a, we notice that the set of images in mOSCAR documents are more diverse than images from WIT documents but less diverse than LAION-400M.

Multilingual diversity We also compare the diversity of images from English documents and of images sampled from documents of any language (English included). We use multilingual SigLIP (Chen et al., 2023a) trained on WebLI (Chen et al., 2023b) to compute image embeddings used to get the Vendi score. We again use a batch of size 65,536 and a total of 3M images, and we do not sample multiple images from a same document. For the multilingual setting, we randomly sample 50 languages and an equal number of images for each language to build the batch. As we did not do any image deduplication across languages, we could expect to have less diversity in the multilingual setting. However, Table 3b shows that the set of images is on average more diverse when sampled from all documents than from English-only documents. This means that the distribution of images is not exactly the same across languages, potentially due to cultural differences.

5 Training a multilingual multimodal language model

We train a multilingual Flamingo-like model on mOSCAR that we call multilingual Open Flamingo. As adding captioning data to training data has been shown to improve zero-shot performance (McKinzie et al., 2024), we additionally train on LAION-400M, which we re-captioned using BLIP (Li et al., 2022), filtered with CLIP score (Radford et al., 2021) and translated using distilled NLLB-600M (Visheratin, 2023) following the proportion of languages found in mOSCAR. We use Gemma-2B (Gemma Team et al., 2024) as the underlying language model and we train the model on 50M mOSCAR documents and 100M randomly sampled image-text pairs. We also train a model on 300M image-text pairs, a model trained on 35M WIT (Srinivasan et al., 2021) documents and 70M text-image pairs and a model trained on 50M mOSCAR documents from the English subset and 100M English image-text pairs as comparison baselines. We additionally compare with OpenFlamingo-3B-MPT (Awadalla et al., 2023) as the *translate-test* baseline. The full list of languages for training and the implementation details can be found in the Supplementary Material.

5.1 Evaluation setup

We evaluate the models using a broad set of image-text multilingual tasks and benchmarks. We use the IGLUE benchmark (Bugliarello et al., 2022) composed of XVNLI, MaRVL (Liu et al., 2021) to test reasoning, xGQA (Pfeiffer et al., 2022) to test visual question answering capabilities and xFlickr&CO (Young et al., 2014; Karpathy and Fei-Fei, 2015; Yoshikawa et al., 2017) for captioning. We also include Crossmodal-3600 (XM3600) (Thapliyal et al., 2022) and MaXM (Changpinyo et al., 2022) as they cover a broader range of languages. To test to what extent models trained on mOSCAR can perform zero-shot multimodal machine translation (MMT), we also test on Multi30K (Elliott et al., 2016; 2017; Barrault et al., 2018) and CoMMuTE (Futeral et al., 2023). For captioning we compute the CideR (Vedantam et al., 2015) score and we tokenize references and model outputs with the Stanford Core NLP tokenizer for English and Stanza (Qi et al., 2020) tokenizers for other languages. To evaluate Multi30k, we compute BLEU (Papineni et al., 2002) score from Sacrebleu (Post, 2018) with *13a* tokenization and default parameters. We use accuracy for CoMMuTE. More details can be found in the Supplementary Material.

5.2 Results

Tables 4 and 6 show the average results across all languages. Full results are available in the Supplementary Material. We notice that the multilingual OpenFlamingo trained additionally on mOSCAR gets better results than the model trained on captioning data only while having seen fewer image-text pairs during training. More importantly, when increasing the number of few-shot examples from 0 to 16, it sees gains of on average +6.71 points on VQA benchmarks and +19.39 CideR points on captioning benchmarks. In contrast, the model trained on text-image pairs only sees gains of +2.82 and +9.08 points respectively. In cross-modal machine translation, the model additionally trained on interleaved data is again much better than the one trained on just captioning data, which is not able to translate the Multi30k benchmark at all.[16] Moreover, mOSCAR helps the model to learn to zero-shot

[16]Most of the time, the model is not able to follow the prompt and only outputs the end of sequence token.

disambiguate translations as shown by the improved average score on CoMMuTE (63.54) compared to the model trained on captions only (61.36).

Multilingual Open Flamingo trained on mOSCAR & text-image pairs is also better than Open-Flamingo 3B MPT evaluated on translate test benchmarks[17]. However, we obtain the best results (except for MaXM) by evaluating our multilingual Open Flamingo on the translate-test benchmarks since the underlying language model (Gemma-2B) is far better in English than other languages. We also notice that all models struggle with reasoning classification tasks (MaRVL, XVNLI) where they obtain scores close to random guessing.

Table 5 additionally shows that Multilingual Open Flamingo trained on mOSCAR obtains much better results than the same model trained on WIT for equivalent training data seen during training[18] (except for Multi30K benchmark) which means mOSCAR is better suited than WIT for training multilingual mLLMs. Eventually, Table 7 shows that we don't face a drop in performances in English performances when training the model on 43 languages (multilingual Open Flamingo) in comparison to training it on the English subset of mOSCAR and English text-image pairs.

Additional comparison results with InternVL2 (Chen et al., 2024), Llava-NeXT (Li et al., 2024), PaliGemma (Beyer* et al., 2024) and Idefics2 (Laurençon et al., 2024) can be found in the Supplementary Material.

	#shots	xFlickR&CO	XM3600	xGQA	MaXM	MaRVL	XVNLI	Multi30K	CoMMuTE
Multi. OF *mOSCAR + cap.*	0	16.91	7.45	26.95	22.33	49.56	33.88	22.91	63.34
	4	34.80	22.18	32.33	26.33	49.64	34.07	23.27	63.22
	8	36.90	23.48	34.24	27.08	**51.48**	**36.60**	23.59	**63.54**
	16	**39.46**	**23.67**	**35.23**	**27.47**	49.84	34.85	**23.85**	62.78
Multi. OF *cap. only*	0	9.57	4.21	8.62	4.01	**49.88**	33.76	0.00	61.36
	4	13.20	9.26	13.45	4.15	49.54	32.04	0.00	61.13
	8	18.00	10.35	12.82	4.88	49.65	33.71	0.01	60.90
	16	19.87	12.07	13.37	4.89	49.79	32.70	0.74	60.25

Table 4: Results averaged over all languages. Multi. OF refers to multilingual Open Flamingo, *mOSCAR + cap.* refers to the model trained on text-image pairs and mOSCAR while *cap. only* refers to the model trained only on text-image pairs. **Bold** is best result.

	#shots	xFlickR&CO	XM3600	xGQA	MaXM	MaRVL	XVNLI	Multi30K	CoMMuTE
Multi. OF (35M) *mOSCAR + cap.*	0	19.07	8.73	25.08	19.64	**49.77**	33.01	22.70	**63.75**
	4	34.32	20.59	31.90	23.90	49.67	36.07	22.79	63.65
	8	36.77	22.15	33.9	24.41	49.72	**37.16**	23.21	63.00
	16	**37.63**	**22.24**	**35.71**	**25.38**	49.73	35.36	23.48	62.77
Multi. OF (35M) *WIT + cap.*	0	9.39	4.67	19.81	14.63	49.71	32.78	**26.99**	56.75
	4	7.68	2.99	25.68	16.12	49.72	33.51	**26.99**	53.27
	8	8.91	3.63	27.06	16.81	49.74	32.77	**26.99**	55.33
	16	9.74	4.14	28.14	16.34	49.74	33.63	**26.99**	54.04

Table 5: Results averaged over all languages and comparison between a model trained on WIT and a checkpoint of multilingual Open Flamingo trained on 35M mOSCAR documents (full model was trained on 50M mOSCAR documents). Both models were trained on 35M documents from their respective training datasets and 70M text-image pairs for fair comparison. Multi. OF (35M) refers to multilingual Open Flamingo trained on 35M documents. **Bold** is best result.

We additionally compare results at different training steps, defined by the number of images seen during training. Figure 4 shows the difference of averaged scores between the model trained on all data and the model trained only on text-images pairs. We notice that the gap first decreases until 20M images seen and keep increasing over training at all training steps after that. Particularly, the gap is wider for few-shot learning.

[17]This means benchmarks were translated from local languages to English, using Google Translate API

[18]We select the checkpoint of multilingual Open Flamingo trained on 35M documents and 70M captions to have fair comparison.

	#shots	xGQA	MaXM	MaRVL	XVNLI
OF-3B MPT	0	18.34	7.68	49.75	32.73
	4	22.97	7.82	49.70	35.82
	8	28.57	8.32	49.71	31.29
	16	31.82	9.04	49.72	33.29
Multi. OF *mOSCAR + cap.*	0	30.16	10.06	49.93	34.66
	4	35.55	9.89	48.99	36.10
	8	36.78	10.12	**50.54**	**39.69**
	16	**37.75**	**11.49**	49.57	37.97

Table 6: *Translate-test* results averaged over languages where all benchmarks were translated from local languages into English using Google Translate API. Multi. OF *mOSCAR + cap.* refers to Multilingual Open Flamingo trained on mOSCAR and text-image pairs while OF-3B MPT refers to Open Flamingo (Awadalla et al., 2023) based on MPT (Team, 2023) and trained on mmc4 (Zhu et al., 2023) and English text-image pairs.

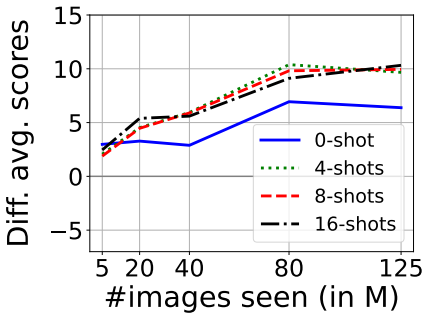


Figure 4: Score differences averaged over benchmarks and languages between the model trained on mOSCAR + text-image pairs and the model trained only on text-image pairs. **Bold** is best result.

	#shots	xFlickR&CO	XM3600	xGQA	MaXM	XVNLI
Multilingual OF *mOSCAR + cap.*	0	29.64	42.57	34.24	36.58	34.62
	4	51.47	77.98	37.91	38.13	33.59
	8	56.75	77.64	39.44	**38.52**	**38.75**
	16	**59.89**	78.18	40.09	35.80	36.60
English OF *English mOSCAR + English cap.*	0	32.70	43.75	34.71	36.19	35.82
	4	51.39	75.33	37.48	37.96	34.88
	8	51.44	77.73	39.64	38.35	36.86
	16	59.24	**78.38**	**40.36**	37.35	37.11

Table 7: Results on the English subsets of the test sets and comparison between multilingual Open Flamingo and an Open Flamingo trained on the English subset of mOSCAR and English text-image pairs (English OF). Both models were trained on 50M documents from their respective training datasets and 100M text-image pairs for fair comparison. **Bold** is best result.

6 Conclusion, Limitations and Societal Impacts

We introduce mOSCAR, a large-scale multilingual and multimodal dataset covering 163 languages and composed of 303M documents, 200B tokens and 1.15B images. We show that mOSCAR is of good quality, diverse and can be used to train a multilingual and multimodal LLM. We ensure that mOSCAR is as safe as possible by applying a series of filtering steps to remove NSFW and toxic content. We however did not conduct any analysis of its biases as this is challenging in a multilingual setting. As it is crawled from the internet, it is indeed possible that mOSCAR reflects biases widespread on it. Training a model on mOSCAR must therefore be combined with additional alignment training steps to mitigate potential biases towards groups of people. Nevertheless, by its multilingual nature, mOSCAR is a step towards the inclusion of more languages, cultures and people in accessing mLLMs.

References

Julien Abadji, Pedro Javier Ortiz Suárez, Laurent Romary, and Benoît Sagot. Ungoliant: An optimized pipeline for the generation of a very large-scale multilingual web corpus. In Harald Lüngen, Marc Kupietz, Piotr Bański, Adrien Barbaresi, Simon Clematide, and Ines Pisetta, editors, *Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-9)*, pages 1–9, Limerick, 2021. Leibniz-Institut für Deutsche Sprache. doi: 10.14618/ids-pub-10468. URL https://nbn-resolving.org/urn:nbn:de:bsz:mh39-104688.

Julien Abadji, Pedro Ortiz Suarez, Laurent Romary, and Benoît Sagot. Towards a cleaner document-oriented multilingual crawled corpus. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache,

Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4344–4355, Marseille, France, June 2022. European Language Resources Association. URL https://aclanthology.org/2022.lrec-1.463.

Amro Abbas, Kushal Tirumala, Dániel Simig, Surya Ganguli, and Ari S Morcos. Semdedup: Data-efficient learning at web-scale through semantic deduplication. *arXiv preprint arXiv:2303.09540*, 2023.

Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems*, 35:23716–23736, 2022.

Anas Awadalla, Irena Gao, Josh Gardner, Jack Hessel, Yusuf Hanafy, Wanrong Zhu, Kalyani Marathe, Yonatan Bitton, Samir Gadre, Shiori Sagawa, Jenia Jitsev, Simon Kornblith, Pang Wei Koh, Gabriel Ilharco, Mitchell Wortsman, and Ludwig Schmidt. Openflamingo: An open-source framework for training large autoregressive vision-language models. *arXiv preprint arXiv:2308.01390*, 2023.

Loïc Barrault, Fethi Bougares, Lucia Specia, Chiraag Lala, Desmond Elliott, and Stella Frank. Findings of the third shared task on multimodal machine translation. In *Proceedings of the Third Conference on Machine Translation: Shared Task Papers*, pages 304–323, 2018.

Romain Beaumont. img2dataset: Easily turn large sets of image urls to an image dataset. https://github.com/rom1504/img2dataset, 2021.

Janek Bevendorff, Benno Stein, Matthias Hagen, and Martin Potthast. Elastic ChatNoir: Search Engine for the ClueWeb and the Common Crawl. In Leif Azzopardi, Allan Hanbury, Gabriella Pasi, and Benjamin Piwowarski, editors, *Advances in Information Retrieval. 40th European Conference on IR Research (ECIR 2018)*, Lecture Notes in Computer Science, Berlin Heidelberg New York, 2018. Springer.

Janek Bevendorff, Martin Potthast, and Benno Stein. Fastwarc: optimizing large-scale web archive analytics. *arXiv preprint arXiv:2112.03103*, 2021.

Lucas Beyer*, Andreas Steiner*, André Susano Pinto*, Alexander Kolesnikov*, Xiao Wang*, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, Thomas Unterthiner, Daniel Keysers, Skanda Koppula, Fangyu Liu, Adam Grycner, Alexey Gritsenko, Neil Houlsby, Manoj Kumar, Keran Rong, Julian Eisenschlos, Rishabh Kabra, Matthias Bauer, Matko Bošnjak, Xi Chen, Matthias Minderer, Paul Voigtlaender, Ioana Bica, Ivana Balazevic, Joan Puigcerver, Pinelopi Papalampidi, Olivier Henaff, Xi Xiong, Radu Soricut, Jeremiah Harmsen, and Xiaohua Zhai*. PaliGemma: A versatile 3B VLM for transfer. *arXiv preprint arXiv:2407.07726*, 2024.

A.Z. Broder. On the resemblance and containment of documents. In *Proceedings of the Compression and Complexity of Sequences 1997*, pages 21–29. IEEE Computer Society, 1997. doi: 10.1109/SEQUEN.1997.666900.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

Emanuele Bugliarello, Fangyu Liu, Jonas Pfeiffer, Siva Reddy, Desmond Elliott, Edoardo Maria Ponti, and Ivan Vulić. Iglue: A benchmark for transfer learning across modalities, tasks, and languages. In *International Conference on Machine Learning*, pages 2370–2392. PMLR, 2022.

Laurie Burchell, Alexandra Birch, Nikolay Bogoychev, and Kenneth Heafield. An open dataset and model for language identification. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 865–879, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-short.75. URL https://aclanthology.org/2023.acl-short.75.

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

Nicholas Carlini, Jamie Hayes, Milad Nasr, Matthew Jagielski, Vikash Sehwag, Florian Tramèr, Borja Balle, Daphne Ippolito, and Eric Wallace. Extracting training data from diffusion models. In *Proceedings of the 32nd USENIX Conference on Security Symposium*, SEC '23, USA, 2023. USENIX Association. ISBN 978-1-939133-37-3.

Nicolas Carlini, Matthew Jagielski, Christopher Choquette-Choo, Daniel Paleka, Will Pearce, Hyrum Anderson, Andreas Terzis, Kurt Thomas, and Florian Tramèr. Poisoning web-scale training datasets is practical. In *Proceedings of the 2024 IEEE Symposium on Security and Privacy (SP)*, pages 179–179, Los Alamitos, CA, USA, 2024. IEEE Computer Society. doi: 10.1109/SP54263.2024.00179. URL https://doi.ieeecomputersociety.org/10.1109/SP54263.2024.00179.

Soravit Changpinyo, Piyush Sharma, Nan Ding, and Radu Soricut. 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, Nashville, TN, USA, 2021.

Soravit Changpinyo, Linting Xue, Idan Szpektor, Ashish V Thapliyal, Julien Amelot, Michal Yarom, Xi Chen, and Radu Soricut. Maxm: Towards multilingual visual question answering. *arXiv preprint arXiv:2209.05401*, 2022.

Xi Chen, Xiao Wang, Lucas Beyer, Alexander Kolesnikov, Jialin Wu, Paul Voigtlaender, Basil Mustafa, Sebastian Goodman, Ibrahim Alabdulmohsin, Piotr Padlewski, et al. Pali-3 vision language models: Smaller, faster, stronger. *arXiv preprint arXiv:2310.09199*, 2023a.

Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, et al. Pali: A jointly-scaled multilingual language-image model. In *Proceedings of the International Conference on Learning Representations*, Kigali, Rwanda, 2023b.

Zhe Chen, Weiyun Wang, Hao Tian, Shenglong Ye, Zhangwei Gao, Erfei Cui, Wenwen Tong, Kongzhi Hu, Jiapeng Luo, Zheng Ma, et al. How far are we to gpt-4v? closing the gap to commercial multimodal models with open-source suites. *arXiv preprint arXiv:2404.16821*, 2024.

Marta R Costa-jussà, James Cross, Onur Çelebi, Maha Elbayad, Kenneth Heafield, Kevin Heffernan, Elahe Kalbassi, Janice Lam, Daniel Licht, Jean Maillard, et al. No language left behind: Scaling human-centered machine translation. *arXiv preprint arXiv:2207.04672*, 2022.

Jesse Dodge, Maarten Sap, Ana Marasović, William Agnew, Gabriel Ilharco, Dirk Groeneveld, Margaret Mitchell, and Matt Gardner. Documenting large webtext corpora: A case study on the colossal clean crawled corpus. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1286–1305, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.98. URL https://aclanthology.org/2021.emnlp-main.98.

Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. Multi30k: Multilingual english-german image descriptions. In *Proceedings of the 5th Workshop on Vision and Language*, pages 70–74. Association for Computational Linguistics, 2016. doi: 10.18653/v1/W16-3210. URL http://www.aclweb.org/anthology/W16-3210.

Desmond Elliott, Stella Frank, Loïc Barrault, Fethi Bougares, and Lucia Specia. Findings of the second shared task on multimodal machine translation and multilingual image description. In *Proceedings of the Second Conference on Machine Translation, Volume 2: Shared Task Papers*, pages 215–233, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. URL http://www.aclweb.org/anthology/W17-4718.

Dan Friedman and Adji Bousso Dieng. The vendi score: A diversity evaluation metric for machine learning. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=g97OHbQyk1.

Matthieu Futeral, Cordelia Schmid, Ivan Laptev, Benoît Sagot, and Rachel Bawden. Tackling ambiguity with images: Improved multimodal machine translation and contrastive evaluation. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5394–5413, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.acl-long.295. URL https://aclanthology.org/2023.acl-long.295.

Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, et al. Datacomp: In search of the next generation of multimodal datasets. *Advances in Neural Information Processing Systems*, 36, 2024.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, et al. The pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*, 2020.

Tianyu Gao, Xingcheng Yao, and Danqi Chen. SimCSE: Simple contrastive learning of sentence embeddings. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 6894–6910, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.552. URL https://aclanthology.org/2021.emnlp-main.552.

Aristides Gionis, Piotr Indyk, Rajeev Motwani, et al. Similarity search in high dimensions via hashing. In *Proceedings of the 25th VLDB Conference*, volume 99, pages 518–529, Edinburgh, Scotland, UK, 1999.

John Hewitt, Daphne Ippolito, Brendan Callahan, Reno Kriz, Derry Tanti Wijaya, and Chris Callison-Burch. Learning translations via images with a massively multilingual image dataset. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Melbourne, Australia, July 2018. Association for Computational Linguistics.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Thomas Hennigan, Eric Noland, Katherine Millican, George van den Driessche, Bogdan Damoc, Aurelia Guy, Simon Osindero, Karén Simonyan, Erich Elsen, Oriol Vinyals, Jack Rae, and Laurent Sifre. Training compute-optimal large language models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 30016–30030. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/c1e2faff6f588870935f114ebe04a3e5-Paper-Conference.pdf.

Andrew G Howard, Menglong Zhu, Bo Chen, Dmitry Kalenichenko, Weijun Wang, Tobias Weyand, Marco Andreetto, and Hartwig Adam. Mobilenets: Efficient convolutional neural networks for mobile vision applications. *arXiv preprint arXiv:1704.04861*, 2017.

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. URL https://doi.org/10.5281/zenodo.5143773.

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

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

Vaclav Kosar, Antonín Hoskovec, Milan Šulc, and Radek Bartyzal. GLAMI-1M: A Multilingual Image-Text Fashion Dataset. In *Proceedings of the 33rd British Machine Vision Conference*, London, UK, 2022. BMVA Press. URL https://bmvc2022.mpi-inf.mpg.de/0607.pdf.

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

Gant Laborde. Deep nn for nsfw detection. URL https://github.com/GantMan/nsfw_model.

Hugo Laurençon, Lucile Saulnier, Thomas Wang, Christopher Akiki, Albert Villanova del Moral, Teven Le Scao, Leandro Von Werra, Chenghao Mou, Eduardo González Ponferrada, Huu Nguyen, Jörg Frohberg, Mario Šaško, Quentin Lhoest, Angelina McMillan-Major, Gerard Dupont, Stella Biderman, Anna Rogers, Loubna Ben allal, Francesco De Toni, Giada Pistilli, Olivier Nguyen, Somaieh Nikpoor, Maraim Masoud, Pierre Colombo, Javier de la Rosa, Paulo Villegas, Tristan Thrush, Shayne Longpre, Sebastian Nagel, Leon Weber, Manuel Muñoz, Jian Zhu, Daniel Van Strien, Zaid Alyafeai, Khalid Almubarak, Minh Chien Vu, Itziar Gonzalez-Dios, Aitor Soroa, Kyle Lo, Manan Dey, Pedro Ortiz Suarez, Aaron Gokaslan, Shamik Bose, David Adelani, Long Phan, Hieu Tran, Ian Yu, Suhas Pai, Jenny Chim, Violette Lepercq, Suzana Ilic, Margaret Mitchell, Sasha Alexandra Luccioni, and Yacine Jernite. The bigscience roots corpus: A 1.6tb composite multilingual dataset. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 31809–31826. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/ce9e92e3de2372a4b93353eb7f3dc0bd-Paper-Datasets_and_Benchmarks.pdf.

Hugo Laurençon, Lucile Saulnier, Leo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander Rush, Douwe Kiela, Matthieu Cord, and Victor Sanh. OBELICS: An Open Web-Scale Filtered Dataset of Interleaved Image-Text Documents. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 71683–71702. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/e2cfb719f58585f779d0a4f9f07bd618-Paper-Datasets_and_Benchmarks.pdf.

Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh. What matters when building vision-language models? *arXiv preprint arXiv:2405.02246*, 2024.

Colin Leong, Joshua Nemecek, Jacob Mansdorfer, Anna Filighera, Abraham Owodunni, and Daniel Whitenack. Bloom library: Multimodal datasets in 300+ languages for a variety of downstream tasks. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8608–8621, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.590. URL https://aclanthology.org/2022.emnlp-main.590.

Vladimir Iosifovich Levenshtein. Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady*, 10(8):707–710, 1966. Doklady Akademii Nauk SSSR, V163 No4 845-848 1965.

Bo Li, Kaichen Zhang, Hao Zhang, Dong Guo, Renrui Zhang, Feng Li, Yuanhan Zhang, Ziwei Liu, and Chunyuan Li. Llava-next: Stronger llms supercharge multimodal capabilities in the wild, May 2024. URL https://llava-vl.github.io/blog/2024-05-10-llava-next-stronger-llms/.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. A diversity-promoting objective function for neural conversation models. In Kevin Knight, Ani Nenkova, and Owen Rambow, editors, *Proceedings of the 2016 Conference of the North American Chapter of the*

Association for Computational Linguistics: Human Language Technologies, pages 110–119, San Diego, California, June 2016. Association for Computational Linguistics. doi: 10.18653/v1/N16-1014. URL https://aclanthology.org/N16-1014.

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

Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. BLIP: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *Proceedings of the 39 th International Conference on Machine Learning*, pages 12888–12900, Baltimore, Maryland, USA, 2022. PMLR.

Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *Proceedings of the 40th International Conference on Machine Learning*, pages 19730—-19742, Honolulu Hawaii USA, 2023.

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 *Proceedings of the 13th European Conference on Computer Vision*, pages 740–755, Zurich, Switzerland, 2014. Springer.

Fangyu Liu, Emanuele Bugliarello, Edoardo Maria Ponti, Siva Reddy, Nigel Collier, and Desmond Elliott. Visually grounded reasoning across languages and cultures. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 10467–10485, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.818. URL https://aclanthology.org/2021.emnlp-main.818.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and Luke Zettlemoyer. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742, 2020. doi: 10.1162/tacl_a_00343. URL https://aclanthology.org/2020.tacl-1.47.

Muhammad Maaz, Hanoona Rasheed, Abdelrahman Shaker, Salman Khan, Hisham Cholakal, Rao M Anwer, Tim Baldwin, Michael Felsberg, and Fahad S Khan. PALO: A Polyglot Large Multimodal Model for 5B People. *arXiv preprint arXiv:2402.14818*, 2024.

Chameleon Team. Chameleon: Mixed-modal early-fusion foundation models. *arXiv preprint arXiv:2405.09818*, 2024.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Gemma Team, Thomas Mesnard, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, et al. Gemma: Open models based on Gemini research and technology. *arXiv preprint arXiv:2403.08295*, 2024.

Brandon McKinzie, Zhe Gan, Jean-Philippe Fauconnier, Sam Dodge, Bowen Zhang, Philipp Dufter, Dhruti Shah, Xianzhi Du, Futang Peng, Floris Weers, et al. MM1: Methods, Analysis & Insights from Multimodal LLM Pre-training. *arXiv preprint arXiv:2403.09611*, 2024.

Owais Khan Mohammed, Kriti Aggarwal, Qiang Liu, Saksham Singhal, Johan Bjorck, and Subhojit Som. Bootstrapping a high quality multilingual multimodal dataset for Bletchley. In Emtiyaz Khan and Mehmet Gonen, editors, *Proceedings of The 14th Asian Conference on Machine Learning*, volume 189 of *Proceedings of Machine Learning Research*, pages 738–753. PMLR, 12–14 Dec 2023. URL https://proceedings.mlr.press/v189/mohammed23a.html.

OpenAI. GPT-4 Technical Report. *ArXiv*, abs/2303.08774, 2023. URL https://arxiv.org/abs/2303.08774.

Vicente Ordonez, Girish Kulkarni, and Tamara Berg. Im2Text: Describing Images Using 1 Million Captioned Photographs. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K.Q. Weinberger, editors, *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011. URL https://proceedings.neurips.cc/paper_files/paper/2011/file/5dd9db5e033da9c6fb5ba83c7a7ebea9-Paper.pdf.

Pedro Javier Ortiz Suárez, Benoît Sagot, and Laurent Romary. Asynchronous pipelines for processing huge corpora on medium to low resource infrastructures. In Piotr Bański, Adrien Barbaresi, Hanno Biber, Evelyn Breiteneder, Simon Clematide, Marc Kupietz, Harald Lüngen, and Caroline Iliadi, editors, *Proceedings of the Workshop on Challenges in the Management of Large Corpora (CMLC-7)*, pages 9 – 16, Cardiff, UK, 2019. Leibniz-Institut für Deutsche Sprache. doi: 10.14618/ids-pub-9021. URL http://nbn-resolving.de/urn:nbn:de:bsz:mh39-90215.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In Pierre Isabelle, Eugene Charniak, and Dekang Lin, editors, *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA, July 2002. Association for Computational Linguistics. doi: 10.3115/1073083.1073135. URL https://aclanthology.org/P02-1040.

Guilherme Penedo, Quentin Malartic, Daniel Hesslow, Ruxandra Cojocaru, Hamza Alobeidli, Alessandro Cappelli, Baptiste Pannier, Ebtesam Almazrouei, and Julien Launay. The refinedweb dataset for falcon llm: Outperforming curated corpora with web data only. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine, editors, *Advances in Neural Information Processing Systems*, volume 36, pages 79155–79172. Curran Associates, Inc., 2023. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/fa3ed726cc5073b9c31e3e49a807789c-Paper-Datasets_and_Benchmarks.pdf.

Jonas Pfeiffer, Gregor Geigle, Aishwarya Kamath, Jan-Martin Steitz, Stefan Roth, Ivan Vulić, and Iryna Gurevych. xGQA: Cross-lingual visual question answering. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2497–2511, Dublin, Ireland, May 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.findings-acl.196. URL https://aclanthology.org/2022.findings-acl.196.

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

Matt Post. A call for clarity in reporting BLEU scores. In Ondřej Bojar, Rajen Chatterjee, Christian Federmann, Mark Fishel, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, Christof Monz, Matteo Negri, Aurélie Névéol, Mariana Neves, Matt Post, Lucia Specia, Marco Turchi, and Karin Verspoor, editors, *Proceedings of the Third Conference on Machine Translation: Research Papers*, pages 186–191, Brussels, Belgium, October 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-6319. URL https://aclanthology.org/W18-6319.

Peng Qi, Yuhao Zhang, Yuhui Zhang, Jason Bolton, and Christopher D. Manning. Stanza: A Python natural language processing toolkit for many human languages. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, 2020. URL https://nlp.stanford.edu/pubs/qi2020stanza.pdf.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *Proceedings of the 38 th International Conference on Machine Learning*, pages 8748–8763, online, 2021. PMLR.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. The risk of racial bias in hate speech detection. In Anna Korhonen, David Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678, Florence, Italy, July 2019. Association for Computational Linguistics. doi: 10.18653/v1/P19-1163. URL https://aclanthology.org/P19-1163.

Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. LAION-400M: Open Dataset of CLIP-Filtered 400 Million Image-Text Pairs. In *Proceedings of the Data Centric AI NeurIPS Workshop 2021*, online, 2021.

Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa 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 S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 25278–25294. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/a1859debfb3b59d094f3504d5ebb6c25-Paper-Datasets_and_Benchmarks.pdf.

Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. Conceptual captions: A cleaned, hypernymed, image alt-text dataset for automatic image captioning. In Iryna Gurevych and Yusuke Miyao, editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2556–2565, Melbourne, Australia, July 2018. Association for Computational Linguistics. doi: 10.18653/v1/P18-1238. URL https://aclanthology.org/P18-1238.

Shaden Smith, Mostofa Patwary, Brandon Norick, Patrick LeGresley, Samyam Rajbhandari, Jared Casper, Zhun Liu, Shrimai Prabhumoye, George Zerveas, Vijay Korthikanti, et al. Using DeepSpeed and Megatron to Train Megatron-Turing NLG 530B, A Large-Scale Generative Language Model. *arXiv preprint arXiv:2201.11990*, 2022.

Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Diffusion art or digital forgery? investigating data replication in diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6048–6058, 2023.

Gowthami Somepalli, Vasu Singla, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Understanding and mitigating copying in diffusion models. *Advances in Neural Information Processing Systems*, 36, 2024.

Krishna Srinivasan, Karthik Raman, Jiecao Chen, Michael Bendersky, and Marc Najork. 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, online, 2021.

Dídac Surís, Dave Epstein, and Carl Vondrick. Globetrotter: Connecting languages by connecting images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16474–16484, Canada, 2022.

MosaicML NLP Team. Introducing mpt-7b: A new standard for open-source, commercially usable llms, 2023. URL www.mosaicml.com/blog/mpt-7b. Accessed: 2023-05-05.

Guy Tevet and Jonathan Berant. Evaluating the evaluation of diversity in natural language generation. In Paola Merlo, Jorg Tiedemann, and Reut Tsarfaty, editors, *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 326–346, Online, April 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.eacl-main.25. URL https://aclanthology.org/2021.eacl-main.25.

Ashish V. Thapliyal, Jordi Pont Tuset, Xi Chen, and Radu Soricut. Crossmodal-3600: A massively multilingual multimodal evaluation dataset. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 715–729, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.45. URL https://aclanthology.org/2022.emnlp-main.45.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. LLaMA: Open and Efficient Foundation Language Models. *arXiv preprint arXiv:2302.13971*, 2023a.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. Llama 2: Open foundation and fine-tuned chat models. *arXiv preprint arXiv:2307.09288*, 2023b.

Ramakrishna Vedantam, C Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4566–4575, 2015.

Alexander Visheratin. NLLB-CLIP – train performant multilingual image retrieval model on a budget. *arXiv preprint arXiv:2309.01859*, 2023.

Weihan Wang, Qingsong Lv, Wenmeng Yu, Wenyi Hong, Ji Qi, Yan Wang, Junhui Ji, Zhuoyi Yang, Lei Zhao, Xixuan Song, et al. CogVLM: Visual Expert for Pretrained Language Models. *arXiv preprint arXiv:2311.03079*, 2023.

Wenhui Wang, Hangbo Bao, Li Dong, Johan Bjorck, Zhiliang Peng, Qiang Liu, Kriti Aggarwal, Owais Khan Mohammed, Saksham Singhal, Subhojit Som, et al. Image as a foreign language: Beit pretraining for all vision and vision-language tasks. *arXiv preprint arXiv:2208.10442*, 2022.

Ryan Webster, Julien Rabin, Loic Simon, and Frederic Jurie. On the de-duplication of laion-2b. *arXiv preprint arXiv:2303.12733*, 2023.

Guillaume Wenzek, Marie-Anne Lachaux, Alexis Conneau, Vishrav Chaudhary, Francisco Guzmán, Armand Joulin, and Edouard Grave. CCNet: Extracting high quality monolingual datasets from web crawl data. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asuncion Moreno, Jan Odijk, and Stelios Piperidis, editors, *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4003–4012, Marseille, France, May 2020. European Language Resources Association. ISBN 979-10-95546-34-4. URL https://aclanthology.org/2020.lrec-1.494.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online, October 2020. Association for Computational Linguistics. URL https://www.aclweb.org/anthology/2020.emnlp-demos.6.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. mT5: A massively multilingual pre-trained text-to-text transformer. In Kristina Toutanova, Anna Rumshisky, Luke Zettlemoyer, Dilek Hakkani-Tur, Iz Beltagy, Steven Bethard, Ryan Cotterell, Tanmoy Chakraborty, and Yichao Zhou, editors, *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online, June 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.naacl-main.41. URL https://aclanthology.org/2021.naacl-main.41.

Yuya Yoshikawa, Yutaro Shigeto, and Akikazu Takeuchi. STAIR captions: Constructing a large-scale Japanese image caption dataset. In Regina Barzilay and Min-Yen Kan, editors, *Proceedings of the*

55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 417–421, Vancouver, Canada, July 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-2066. URL https://aclanthology.org/P17-2066.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 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, 2014. doi: 10.1162/tacl_a_00166. URL https://aclanthology.org/Q14-1006.

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

Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. Sigmoid loss for language image pre-training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 11975–11986, October 2023.

Wanrong Zhu, Jack Hessel, Anas Awadalla, Samir Yitzhak Gadre, Jesse Dodge, Alex Fang, Youngjae Yu, Ludwig Schmidt, William Yang Wang, and Yejin Choi. Multimodal c4: An open, billion-scale corpus of images interleaved with text. *arXiv preprint arXiv:2304.06939*, 2023.