# LAION-5B: An open large-scale dataset for training next generation image-text models

**Christoph Schuhmann**[1] §§°° **Romain Beaumont**[1] §§°° **Richard Vencu**[1,3] §§°°
**Cade Gordon**[2] §§°° **Ross Wightman**[1]§§ **Mehdi Cherti** [1,7]§§
**Theo Coombes**[1] **Aarush Katta**[1] **Clayton Mullis**[1]
**Patrick Schramowski**[1,4] **Srivatsa Kundurthy**[1] **Katherine Crowson**[1]
**Ludwig Schmidt**[6] **Robert Kaczmarczyk**[1,5] °° **Jenia Jitsev**[1,7] °°
LAION[1] UC Berkeley[2] Gentec Data[3] Technical University Darmstadt[4]
Technical University of Munich[5] University of Washington, Seattle[6]
Juelich Supercomputing Center (JSC), Research Center Juelich (FZJ) [7]
`contact@laion.ai`
`§§ Equal first contributions, °° Equal senior contributions`

## Abstract

Groundbreaking language-vision architectures like CLIP and DALL-E proved the utility of training on large amounts of noisy image-text data, without relying on expensive accurate labels used in standard vision unimodal supervised learning. The resulting models showed capabilities of strong out-of-distribution sample generation and transfer to downstream tasks, while performing remarkably at zero-shot classification with noteworthy out-of-distribution robustness. Since then, further large-scale language-vision models like ALIGN, BASIC, GLIDE, Flamingo and Imagen made further improvements. Studying the capabilities of such models requires datasets containing billions of image-text pairs. Until now, no datasets of this size have been made openly available for the broader research community. To address this problem and democratize research on large-scale multi-modal models, we present LAION-5B - a dataset consisting of 5.85 billion CLIP-filtered image-text pairs, of which 2.32B contain English language. We show successful replication and fine-tuning of foundational models like CLIP and GLIDE using the dataset, and discuss further experiments enabled with an openly available dataset of this scale. Additionally we provide several nearest neighbor indices, an improved web-interface for exploration and subset generation, and detection scores for watermark, NSFW, and toxic content detection. [1]

## 1 Intro

Learning from multimodal data is a longstanding research challenge in machine learning. Recently, contrastive loss functions combined with large neural networks have achieved significant performance leaps by combining vision and language data [34, 35, 39]. Among the resulting generalization capabilities, one prominent example is data efficient zero- and few-shot classification on downstream tasks. For instance, CLIP [34] showed a large gain in zero-shot classification accuracy on ImageNet ILSVRC-2012 [38], going from 11.5% top-1 accuracy [24] to 76.2%. Inspired by CLIP's success, numerous groups have further increased CLIP's generalization ability by scaling compute, batch, and dataset size [17, 32, 59]. Another recent success of multimodal learning is in generative modelling, where DALL-E [35] and successors [36, 39] demonstrated the potential of text-guided image generation and produced high-quality images specific to the provided prompt.

[1] Project page: https://laion.ai/laion-5b-a-new-era-of-open-large-scale-multi-modal-datasets/

Preprint. Under review.

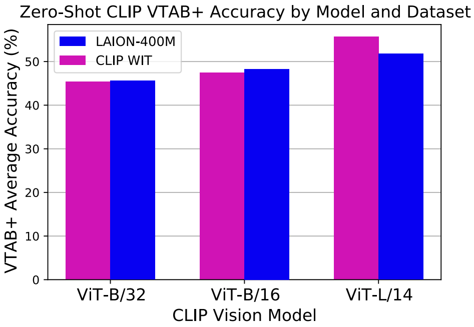

Figure 1: **Zero-Shot Accuracy.** CLIP models trained on LAION-400M [42], a preliminary subset of LAION-5B released previously, show competitive zero-shot accuracy compared to CLIP models trained on OpenAI's original training set WIT when evaluated on the 35 datasets in our new test suite VTAB+.

| Dataset | # of English image-text pairs |
|---|---|
| MS-COCO | 330K |
| CC3M | 3M |
| Visual Genome | 5.4M |
| WIT | 5.5M |
| CC12M | 12M |
| RedCaps | 12M |
| **LAION-5B** | **2.3B** |
| CLIP WIT | 400M |
| ALIGN | 1.8B |
| BASIC | 6.6B |

Table 1: **Dataset Size.** LAION-5B is more than 100 times larger than other public English image-text datasets. Extending RedCaps' analysis [8], we compare public (top) and private (bottom) image-text dataset sizes.

Besides model scale, all of the aforementioned approaches for learning multimodal representations require large volumes of image-text data that exceed hundreds of millions of samples. As the community developed further improvements to multimodal learning with ALIGN, CoCa, Flamingo, and LiT, pre-trained models and large-scale datasets originating from these works were not made available to the broader research community [1, 17, 55, 59]. As such, research in this area, at this scale, has pooled into only a handful of institutions. In this work we take an important step forward by democratizing such training procedures, making available data at sufficiently large scale and publishing pre-trained models originating from the training.

We present LAION-5B, the largest public image-text dataset of over 5.8 billion pairs. Through a grass-roots effort to scrape Common Crawl, we derived 2.32 billion english, 2.26 billion multilingual, and 1.27 billion salient but non-concrete language samples through CLIP-based filtering. We also explore the ethical implications and flaws that emerge with large scale data curation. By releasing the data publicly, in turn, we offer the first opportunity for the community to audit and refine a dataset of this magnitude. Our main contributions are as follows:

- We establish the largest community-available image-text pair dataset.
- We provide exact software and a reproducible pipeline for the data curation process, enabling community uptake and improvement.
- We release the first open-source reproduction of CLIP models up to ViT L/14 scale achieving competitive zero-shot classification and retrieval.

LAION-5B is *not* a finished data product. Due to the immense size of current image-text pre-training datasets, curating LAION-5B for widespread use goes beyond the scope of a single research paper. We view our initial data release and this paper as a first step on the way towards a widely applicable pre-training dataset for multimodal models. As a result, **we strongly recommend that LAION-5B should only be used for academic research purposes in its current form, and we advise against any applications in deployed systems without careful investigation of biases arising in models trained on LAION-5B.**

The remainder of the paper proceeds as follows. First, we present the data collection methodology we use to assemble LAION-5B. Then, we explicitly describe LAION-5B's composition including its various subsets. To validate dataset's value, we reproduce or fine-tune different SOTA language-vision models requiring datasets containing billions of image-text pairs and show that training with LAION-5B leads to comparable performance on important metrics as the original datasets not available publicly. Before concluding we discuss the limitations of a Common Crawl based dataset acquisition as well as safety and ethics concerns arising from the procedure.

## 2 Related Works

**Scaling Laws.** Natural language processing (NLP) exemplified the power of scale in the context of model, dataset, and compute size through empirical study and feats like GPT-3. [5, 19] Similarly, computer vision profited from growth along those same axes and advances in both convolutional and transformer architectures. [9, 21, 53, 57]

To facilitate the further exploration of scale, researchers had to curate even larger datasets. Community efforts like the The Pile aimed to democratize access to such datasets within NLP. [11] In vision, large private datasets such as Instagram-1B, JFT300M, and JFT3B emerged to support image pretraining tasks. [28, 49, 58]

**Vision-Language Models.** CLIP marked a large step forward in multimodal progress for image-text representation. [34] The authors proposed a contrastive learning scheme to embed both images and text into a shared space. The joint representation allowed for impressive zero- and few-shot results. In the realm of classification robustness, CLIP offered a consistent way to increase effective robustness. [50, 54] Moreover, the community proved its utility in a variety of vision tasks beyond classification from captioning, object navigation, and visual question and answering. [20, 30, 44]

In addition to image-text representation, groups also proved the possibility of text to image generation. DALL-E was the first work to illustrate the capability. [35] GLIDE, Imagen, and DALL-E2 followed suit improving visual fidelity and text-prompt correspondence. [31, 36, 39]

After CLIP's initial success, works like ALIGN and BASIC improved upon the results by increasing dataset and batch size. [17, 32] Similarly, LiT increased size along those axes, but also proposed to freeze the image encoder. [59] Other techniques have explored the power of caption generation to answer questions and improve representational power. [1, 26, 55]

**Image-Text Datasets.** Initially, efforts like MS-COCO and Visual Genome curated image and region captions through human annotation, but they only achieved 330K and 5M pairs respectively. [22, 27] Works like YFCC100M provided images and videos with metadata, however a caption couldn't always be guaranteed. [51] Researchers followed suit and works like CC3M utilized the alt-text associated with images and performed cleaning procedures. [43] To increase datascale, the team relaxed the initial filtering protocol and was able to arrive at the subsequent CC12M. [7]. Efforts on alt-text collection continued into ALT200M[15] and ALIGN[17]. Authors of the works increased the data scale up to 1.8 Billion image-text pairs through the usage of alt-text. In contrast, RedCaps sought to use the captions provided by Reddit to collect higher quality captions. [8]

## 3 Collection Methodology

LAION-5B is assembled from Common Crawl, a multi-petabyte corpus of data collected over 12 years of web crawling. In the following we introduce an efficient pipeline to collect, automatically filter, and tag a Common Crawl HTML alt-text, vision-language dataset.

### 3.1 Pipeline

**Acquisition.** The acquisition pipeline follows the flowchart of Fig. 2 and can be split into three major components: distributed processing of the petabyte-scale Common Crawl dataset, distributed download of images, and post-processing.

**Distributed processing of Common Crawl.** To create image-text pairs, we parse the HTML `IMG` tags containing an alternative description from Common Crawl's WAT files. At the same time we perform language detection using CLD3 on text with three possible outputs: English, another language, or no language (contains samples below confidence thresholds [42]). From random sample analysis, the no language set appears to contain short form text depicting names and places.

Once extracted, the data is packed and sent to a PostgreSQL node for storage using the COPY command. The PostgreSQL server was maintained to keep about 500M records at all times by means of balancing the ingress and egress of data from the database.

**Distributed downloading of the images.** We download the raw images from the parsed URLs with asynchronous requests using Trio and Asks libraries in order to maximize resource utilization. To

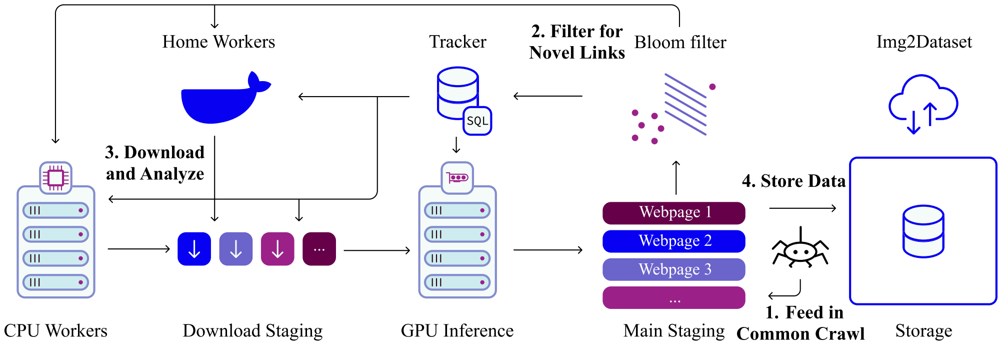

Figure 2: **Overview of the acquisition pipeline:** Files are downloaded, tracked, and undergo distributed inference to determine inclusion. Those above the specified CLIP threshold are saved.

keep costs constrained, we choose a modest cloud node with 2 vCPUs, 1GB RAM, and 10Mbps download bandwidth to use as a worker instance. Such a worker can process 10000 links in about 10-15 minutes; we use roughly 300 workers in parallel. We batch the workload into 10000 link chunks, taken from the PostgreSQL server by using TABLESAMPLE technique to ensure that the distribution among the 10000 links follows that of the 500M records available in the database. With 300 workers, we found that the distribution is good if the database remains above 20M records to be processed. This approach maximized download speed and minimized IP reputation damage.

**Post-Processing.** To conclude the pipeline, we use CLIP's released ViT-B/32 to compute cosine similarities of the English image-text pairings, and M-CLIP ViT-B/32 for all others. Although larger models were later released, they were not available at the time of curation. We remove all English pairs below 0.28 and all others below 0.26, removing around 90% of the original 50 billion images, arriving just short of 6 billion pairs.

### 3.2 Safety During Collection

After downloading the WAT files from Common Crawl, we follow the filtering in Schuhmann et al. [42], focusing on removing data with less than 5 characters of text, less than 5 KB of image data, or containing potentially malicious, large, or redundant images. CLIP embeddings are then computed for the remaining images and are used to filter out illegal content.

Current automated filtering techniques are far from perfect; images that could cause discomfort and disturbance are likely to pass, and others are likely to be falsely removed. We make a best effort to identify, document, and tag such content. Furthermore, these images and texts could amplify the social bias of machine learning models, especially ones trained with no or weak supervision [48].

To encourage research in fields such as dataset curation, we refrain from removing potentially offensive samples and tag them instead. The user can decide whether to include content depending on their task. To this end, we also encourage model developers to state, e.g., in their model card [29] which subsets and tagged images are used.

We apply Q16 [41] and our own specialized pornographic and sexualized content classifier (here referred to as NSFW) to identify and document a broad range of inappropriate concepts displaying not only persons but also objects, symbols, and text, see *cf.* [41] and Appendix Sec. C.5 and Sec. C.6 for details. Both classifiers are based on CLIP embeddings. Following our main intention of a publicly available dataset, these two approaches, as with all other implementations related to LAION 5B, are open-sourced.

We separate pornographic content and otherwise inappropriate content (e.g. harm, exploitation and degradation). Both can be dis- and enabled in the publicly available dataset exploration UI.[2] With both together, the UI and the openly accessible code, we encourage users to explore and, subsequently, report further not yet detected content and thus contribute to the improvement of our and other existing approaches.

[2] https://knn5.laion.ai/

Figure 3: **LAION-5B examples.** Sample images from a nearest neighbor search in LAION-5B using CLIP embeddings. The image and caption (C) are the first results for the query (Q).

It is important to note that the above mentioned classifiers are not perfect, especially keeping the complexity of these tasks and the diverse opinions of different cultures in mind. Interestingly, we observed that the CLIP-based classifiers tend to be more conservative. For instance, samples that seem semantically related to emotional contexts like flirtatious expressions and texts containing sexual words might be tagged as inappropriate. Therefore, we advocate using these tags responsibly, not relying on them to create a truly safe, "production-ready" subset after removing all potentially problematic samples. For a detailed discussion in this regard, we refer to Sec. 7.

## 4 Dataset Composition

We officially release the following subsets of LAION-5B: 2.32 billion of these contain texts in English language, 2.26 billion contain texts from 100+ other languages and 1.27 billion have texts where a particular language could not be clearly detected. We refer to the 2.32 billion subset with English text captions as LAION-2B-en on the remaining text.

We provide parquet files that consist of the following attributes for each pair: sample ID, URL, type of Creative Commons license (if applicable), NSFW tag (detected with CLIP), cosine similarity score between the text and image embedding and height and width of the image. We found 3% of images were detected as NSFW, which can be filtered out by a user with the NSFW tag.

Our multilingual subset contains over 100 languages. The top-5 most frequent languages are Russian (10.6%), French (7.4%), German (6.6%), Spanish (6.6%), and Chinese (6.3%). In the past, researchers had limited outlets to train non-english CLIP-like models. Past researchers opted for translations of english captioning datasets [40, 45, 46], and even when using a multilingual dataset, to the best of our knowledge, had at most 36 million samples from Wikipedia Image Text [47]. With the release of this dataset, multilingual researchers now have access to roughly two orders of magnitude greater samples. This scale provides new opportunities for multilingual and potentially low-resource language researchers.

Upon visual inspection of a random sample of the low-confidence language samples, such images often depict products or places. The captions often contain language with clear semantics, but might include noisy elements like SEO or product tags.

## 5 Experiments Validating the Utility of LAION-5B

In this section, we showcase prior work using the LAION-400M [42] and other subsets as well as our CLIP reproduction studies to give quantitative and qualitative evidence of the dataset's utility for training SOTA large scale language-vision models.

### 5.1 Usage Examples

**Subdataset Generation.** LAION-5B's scale offers potential for curation of datasets for particular computer vision related tasks. Recently, researchers have utilized both LAION-5B and a subset, LAION-400M, as a data source in vision related tasks such as facial representation learning [60] and invasive species mitigation [23]. Within LAION, we have compiled from LAION-5B both

LAION-High-Resolution[3], a 170M subset for superresolution models, and LAION-Aesthetic[4], a 120M subset of aesthetic images, as determined by a linear estimator on top of CLIP.

**CLIP Reproduction.** In Gao et al. [12], an enhanced CLIP architecture was trained on the 400M subset, outperforming OpenAI's CLIP on ImageNet zero-shot classification top-1 accuracy. See Sec. 5.2 for our CLIP reproduction experiments using models of different scales.

**BLIP Training and MAGMA.** Training on a LAION-5B subset, Li et al. [25] developed BLIP to unify understanding and generation for vision-language tasks via a novel Vision-Language Pretraining (VLP) framework. It has been shown that BLIP matched or outperformed comparable models as per CIDEr, SPICE, and BLEU@4 metrics. Eichenberg et al. [10] used a LAION subset for MAGMA, a model generating text "answers" for image-question pairs; MAGMA achieves state of the art results on OKVQA metrics and outperforming *Frozen* [52].

**Image Generation.** Rombach et al. [37] applied a subset of LAION-5B in training Latent Diffusion Models that achieved state-of-the-art results on image inpainting and class-conditional image synthesis. Furthermore, Gu et al. [14] used 400M to train VQ diffusion text-to-image generation models, which have been shown to be more efficient, and are able to generate higher quality images. Moreover, Saharia et al. [39] showed an improved architecture of a diffusion model that was trained on a subset of LAION-400M that even beat OpenAI's recent DALLE-2 and achieved a new state-of-the-art COCO FID of 7.27.

### 5.2 Experiments on CLIP Reproduction

In an effort to reproduce the results of CLIP [34], and to validate the data collection pipeline we describe in Sec. 3, we trained several models on LAION-400M [42] and a model on LAION-2B-en, datasets which are both subsets of LAION-5B. As training such models require large compute due to dataset and model sizes that are considered in the experiments, usage of supercomputers and large machine clusters is necessary in order to train the models efficiently.

We used OpenCLIP [16], an open source software for training CLIP-like models. After adapting OpenCLIP for distributed training and execution on JUWELS Booster supercomputer [18], CLIP models of different size were reproduced on the LAION-400M subset. We trained ViT-B/32, ViT-B/16, and ViT-L/14 following CLIP [34], and an additional model that we call ViT-B/16+, a slightly larger version of ViT-B/16. We followed the same hyper-parameter choices of the original CLIP models. We used between 128 and 400 NVIDIA A100 GPUs to train the models. For more information about hyper-parameters and training details, see Appendix Sec. E.1.

#### 5.2.1 Zero-Shot Classification and Robustness Performance

Following CLIP [34] and subsequent works, we evaluate the models on zero-shot classification. For each downstream dataset, we use a set of pre-defined prompts for each class, which we collected from prior works [34, 59]. We compute the embeddings of each class by averaging over the embedding of the prompts, computed each using the text encoder. For each image, and for each class, we compute the cosine similarity between their embeddings, and classify each image as the class that have the largest cosine similarity with the image embedding. We evaluate the models using top-1 accuracy.

In Tab. 2, we show a comparison between models trained on LAION (400M, 2B) and original CLIP from [34]. We follow [59] and evaluate robustness performance on ImageNet distribution shift datasets. Additionnaly, we construct a benchmark we call VTAB+, a superset of VTAB [56], on which we compute the average top-1 accuracy over 35 tasks[5]. We can see that on ImageNet-1k (noted "INet" on the table), performance of LAION-400M models and original CLIP models (trained on a 400M private dataset) is matched well. On the four ImageNet distribution shift datasets, we observe some larger differences, notably on ObjNet (CLIP WIT better) and INet-S (LAION better), which allows us to conclude that in overall, CLIP models trained on LAION match in their robustness original CLIP. With ViT-B/32, training on the larger LAION-2B-en improves over LAION-400M

[3] https://huggingface.co/datasets/laion/laion-high-resolution

[4] https://github.com/LAION-AI/laion-datasets/blob/main/laion-aesthetic.md

[5] [56] showed that different aggregation strategies have high rank correlation (Kendall score) with the simple top-1 average accuracy over datasets, thus we follow the same strategy. We also compute the ranks of each model on each task and average the ranks, and find that the ranking is similar to averaging top-1 accuracy.

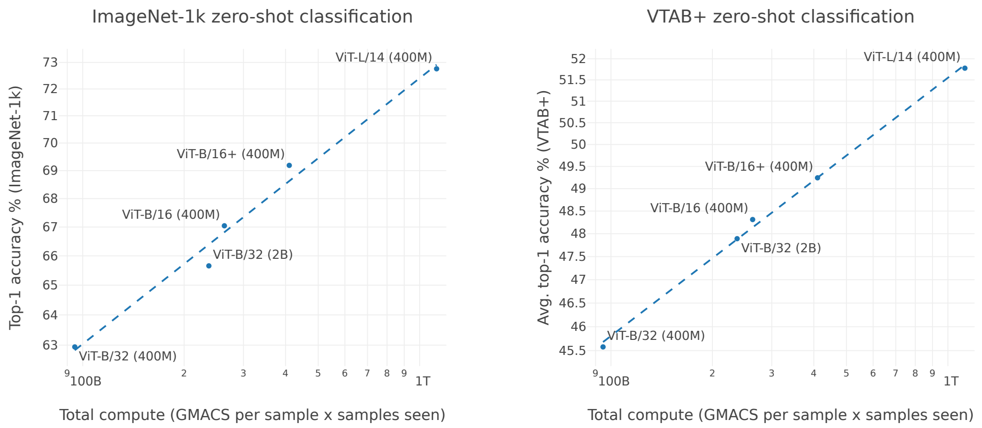

Figure 4: The relationship between total compute (giga multiply–accumulates (GMACS)) and zero-shot top-1 classification accuracy (%) of models trained on LAION (400M, 2B). The dashed line in each figure is a linear fit in log-log space. Each point corresponds to a model trained on either the 400M or 2B LAION subsets. We show results on ImageNet-1k (left) and VTAB+ (right) where we average the accuracy over 35 tasks (see Appendix E.2 for details).

| **Model** | **Pre-training** | INet | INet-v2 | INet-R | INet-S | ObjNet | VTAB+ |
|---|---|---|---|---|---|---|---|
| B/32 | CLIP WIT | 63.3 | 56.0 | 69.4 | 42.3 | 44.2 | 45.4 |
| | LAION-400M | $62.9^{-0.4}$ | $55.1^{-0.9}$ | $73.4^{+4.0}$ | $49.4^{+7.1}$ | $43.9^{-0.3}$ | $45.6^{+0.2}$ |
| | LAION-2B-en | $65.7^{+2.4}$ | $57.4^{+1.4}$ | $75.9^{+6.5}$ | $52.9^{+10.6}$ | $48.7^{+4.5}$ | $47.9^{+2.5}$ |
| B/16 | CLIP WIT | 68.3 | 61.9 | 77.7 | 48.2 | 55.3 | 47.5 |
| | LAION-400M | $67.0^{-1.3}$ | $59.6^{-2.3}$ | $77.9^{+0.2}$ | $52.4^{+4.2}$ | $51.5^{-3.8}$ | $48.3^{+0.8}$ |
| B/16+ | LAION-400M | 69.2 | 61.5 | 80.5 | 54.4 | 53.9 | 49.2 |
| L/14 | CLIP WIT | 75.6 | 69.8 | 87.9 | 59.6 | 69.0 | 55.7 |
| | LAION-400M | $72.8^{-2.8}$ | $65.4^{-4.4}$ | $84.7^{-3.2}$ | 59.6 | $59.9^{-9.1}$ | $51.8^{-3.9}$ |

Table 2: Comparison between CLIP models trained on LAION (400M, 2B) and the original CLIP models [34] trained on OpenAI's WebImageText (WIT) dataset. We show zero-shot top-1 classification accuracy (%) on various datasets including ImageNet, four ImageNet distribution shift datasets, and a benchmark we call VTAB+, where we average performance over 35 tasks. See Appendix E.2 for more details about the datasets used for evaluation and the results.

model everywhere. Overall, on VTAB+, performance of LAION and CLIP WIT models are similar, except on ViT-L/14, where we observe an advantage of CLIP WIT. See Appendix Sec. E.1 for more details about the datasets used for evaluation and the results.

To obtain an idea about how the zero-shot performance improves with scale, we show the relationship between the total compute and accuracy on VTAB+ on models trained on LAION (400M, 2B). In Figure 4, we see that accuracy on VTAB+ improves with compute (log-log plot). It would be interesting to study in future work if the relationship between compute and accuracy keeps showing the same trend or whether we start to see saturation, like it was observed in [58]. Here, we can report that increasing either model or data scale for CLIP pre-training results in improvement of zero-shot classification performance on various downstream transfer targets.

For a full overview of zero-shot classification and retrieval results, view Sec. E.3 of the Appendix.

# 6 Technical Limitations

The large scale of current image-text pre-training datasets such as LAION-5B makes it infeasible to thoroughly investigate all aspects of the dataset construction process. Below we outline some of the potential technical limitations that affect LAION-5B specifically.

**Data Overlap.** Throughout our experiments we find strong results on a variety of downstream tasks. However, it should be noted that data overlap is plausible if Common Crawl contains a certain test-bed. CLIP found few examples of statistically significant performance increases due to data overlap. The same may or may not be true for our analysis and opens the door for future work.

**Usage of Alternative Text.** Birhane et al. [4] began the discussion of the shortcomings by noting that alternative text does not necessarily serve as a caption for the image. At times the text might be Search Engine Optimization (SEO) spam, riddled with keywords, or overly descriptive. In these cases, the language becomes less natural. A surprising result in the case of ImageNet Zero-Shot classification, BASIC [32] succeeded with 5 billion of the 6.6 billion captions being in the form of `CLASS_1 and CLASS_2 and ... and CLASS_K`, i.e. by concatenating class names to form the caption of each image, using an internal multi-label classification dataset (JFT). Such a finding adds nuance to the role of natural language in contrastive image-language models' zero-shot performance.

**Filtering with CLIP.** CLIP allows the curation and collection of this dataset to be low-cost and scalable. With an automated process, one does lose the human control over much of the process. Through curating with CLIP, we also incur its flaws. We further discuss this in subsection 7.1.

Another natural critique is that the dataset won't encapsulate anything beyond CLIP's original knowledge. Naturally if CLIP is presented with a pairing that isn't sufficiently related to its trained distribution, then it shouldn't be given a high cosine similarity. On the other hand, by browsing the KNN indices we find terms that the original CLIP could not have seen such as the United States' "Capitol Insurrection" and "blockage of the Suez Canal." Possible explanations for this are CLIP's optical character recognition abilities and the visual salience of known structures (i.e. it can recognize the United States' Capitol, thus it may include insurrection photos because it recognizes the capitol).

# 7 Safety and Ethical Discussion

Recent developments in large-scale models, such as GPT-3 [6], CLIP [33], ALIGN [17], GLIDE [31] and DALLE-2 [36] have potential for far-reaching impact on society, both positive and negative, if deployed in various end-user settings (e.g., image classification and generation, recommendation systems, search engines). Besides model parameter scaling, the advances made so far also rely on the underlying large-scale datasets. Recent research [2, 3] described many potential negative societal implications that may arise due to careless use of vision-language models trained so far, regarding, e.g., various biases shown towards groups at margins that originate from the composition of training datasets and model training procedure.

Unfortunately, only a minority of these models are publicly released, most of them are only accessible by an "input to output" interface. Importantly, the underlying large-scale datasets are also not often publicly available. Whereas open-source efforts exist to re-implement model architectures and training, the dataset availability issue limits investigations and progress in research to those same institutions that conduct training of large-scale models. Consequently, LAION-5B provides not only a chance to make progress in careful studies of the trained models' capabilities and replication but also to investigate how uncurated large-scale datasets impact various model biases and to design automated ways to curate and create datasets from uncurated ones that alleviate the bias issues. To this end, besides the image-text pairs, we provide pre-computed image embeddings and search functionalities via an easily accessible web-interface.

Indeed, after the release of LAION-400M, several groups (e.g., [4]) investigated potential problems arising from an unfiltered dataset. Motivated by these findings, with LAION-5B, we introduced an improved inappropriate content tagging (*cf.* Sec. 3.2) as well as a Watermark filter, which can improve the quality of the text-to-image models trained on the dataset.

We strongly advocate academic use-only and advise careful investigation of downstream model biases. Additionally, we encourage users to transparently explore and, subsequently, report further not yet detected content to our dataset repository[6] as well as model behaviour, and help to further advance existing approaches for data curation using the real-world large dataset introduced here.

This dataset acts as a starting point, not the final endpoint, for creating models for various tasks. Whereas the vast size of the dataset is the key contribution, its subsets pave the path for future improvements, of which the most obvious is the high-resolution LAION-5B subset of 170M image-text pairs. In recent works, [31] removed images of violence related objects but also images portraying people and faces in order to train a generative model unable to produce e.g. racist content. However, this clearly limits generic capabilities–for instance, generation of human faces–of the model. Therefore, from a functional safety perspective, a potentially more systematic and complete effort to ensure safe usage is the creation of a diverse, balanced, less biased large-scale dataset without restricting content, and tools for extracting subsets that can fulfil required safety constrains suitable for training that aim specific end use scenarios. In our opinion, this process is not supposed to be a closed-door avenue. It should be approached by broad research community, resulting in open and transparent datasets and procedures. Towards meeting this challenge, the large-scale public image-text dataset of over 5.8 billion pairs and further annotations introduced here provides diversity that can be a starting point for ensuring balance and for selecting safe, curated subsets for corresponding target applications. We encourage everybody to participate in this exciting and important future journey.

**Privacy.** Although all photos are publicly available, this large weakly-filtered dataset might pose privacy risks for people who did not expect to appear in online images. Thus, we provide a contact form on our website [7] where removal requests get processed.

### 7.1 CLIP Induced Bias

**Unknown initial dataset.** The CLIP model in itself introduces a bias, which cannot be trivially assessed, as the underlying dataset on which the model was trained is not openly accessible. With the release of a large openly accessible image-text dataset, we offer a starting point in the open auditing of contrastive image-text models like CLIP.

**Selection heuristic based on cosine similarity.** As noted by [4], cosine similarity is only a heuristic that also may lead to suboptimal guidance for dataset filtering. The work showed examples in which captions with malignant descriptions obtain a higher similarity over a benign description. During CLIP's training, the cosine similarity only acted as a logit to represent the likelihood of a given image-text pairing. It fails to encapsulate the nuance and rich semantic and contextual meaning that the image or language might contain. By using cosine similarity as a ground for filtering, the dataset might exacerbate those biases already contained by CLIP.

## 8 Conclusion

By releasing LAION-5B, a larger updated version of an openly available dataset that contains over 5 billion image-text pairs, we have further pushed the scale of open datasets for training and studying state-of-the-art language-vision models. This scale gives strong increases to zero-shot transfer and robustness.

To validate the utility of LAION-5B, we demonstrated that a subset of our dataset can be used to train SOTA CLIP models of various scale that match the strong zero-shot and robustness performance of the original models trained on closed curated data, or to fine-tune generative models like GLIDE, producing samples of good quality. The dataset thus provides opportunities in multi-language large-scale training and research of language-vision models, that were previously restricted to those having access to proprietary large datasets, to the broader research community. Finally, thanks to its large scale, even a rather strict subset filtering (driven by various criterion like NSFW, watermark presence, resolution) provides high-quality datasets that are still large enough to provide sufficient scale for the training or fine-tuning of strong specialized language-vision models.

[6] https://github.com/laion-ai/laion5b-bias

[7] https://laion.ai/#contact

## Acknowledgments

We thank Phil Wang, the creator of the DALLE-pytorch github repository[8], who inspired us and helped creating our open community. We also want to thank Aran Komatsuzaki, Andreas Köpf, Bokai Yu, John David Pressman, Natalie Parde, Mitchell Wortsman, Gabriel Ilharco (see also appendix) and all the members of the LAION discord server [9] for helping crawling image-text-pairs and run inference on their private computers. We want to thank Hugging Face and Stability AI for their continuous financial support and providing hosting space for open datasets and models.

The authors gratefully acknowledge the Gauss Centre for Supercomputing e.V. [10] for funding this work by providing computing time through the John von Neumann Institute for Computing (NIC) on the GCS Supercomputer JUWELS Booster[18] at Jülich Supercomputing Centre (JSC). We also acknowledge storage resources on JUST[13] granted and operated by JSC.

## References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katie Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *arXiv preprint arXiv:2204.14198*, 2022.

[2] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, pages 610–623, 2021.

[3] Abeba Birhane and Vinay Uday Prabhu. Large image datasets: A pyrrhic win for computer vision? In *Proceedings of IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1536–1546. IEEE, 2021.

[4] Abeba Birhane, Vinay Uday Prabhu, and Emmanuel Kahembwe. Multimodal datasets: misogyny, pornography, and malignant stereotypes. October 2021.

[5] 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.

[6] Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. Language models are few-shot learners, 2020. URL `https://arxiv.org/abs/2005.14165`.

[7] 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, 2021.

[8] Karan Desai, Gaurav Kaul, Zubin Aysola, and Justin Johnson. Redcaps: Web-curated image-text data created by the people, for the people. *arXiv preprint arXiv:2111.11431*, 2021.

[9] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[10] Constantin Eichenberg, Sidney Black, Samuel Weinbach, Letitia Parcalabescu, and Anette Frank. MAGMA - multimodal augmentation of generative models through adapter-based finetuning. *CoRR*, abs/2112.05253, 2021. URL `https://arxiv.org/abs/2112.05253`.

[8] https://github.com/lucidrains/DALLE-pytorch

[9] https://discord.gg/xBPBXfcFHd

[10] https://gauss-centre.eu

[11] 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.

[12] Yuting Gao, Jinfeng Liu, Zihan Xu, Jun Zhang, Ke Li, and Chunhua Shen. Pyramidclip: Hierarchical feature alignment for vision-language model pretraining, 2022. URL https://arxiv.org/abs/2204.14095.

[13] Stephan Graf and Olaf Mextorf. Just: Large-scale multi-tier storage infrastructure at the jülich supercomputing centre. *Journal of large-scale research facilities JLSRF*, 7:180, 2021.

[14] Shuyang Gu, Dong Chen, Jianmin Bao, Fang Wen, Bo Zhang, Dongdong Chen, Lu Yuan, and Baining Guo. Vector quantized diffusion model for text-to-image synthesis. *CoRR*, abs/2111.14822, 2021. URL https://arxiv.org/abs/2111.14822.

[15] Xiaowei Hu, Zhe Gan, Jianfeng Wang, Zhengyuan Yang, Zicheng Liu, Yumao Lu, and Lijuan Wang. Scaling up vision-language pre-training for image captioning. *arXiv preprint arXiv:2111.12233*, 2021.

[16] 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. If you use this software, please cite it as below.

[17] 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. *CoRR*, abs/2102.05918, 2021. URL https://arxiv.org/abs/2102.05918.

[18] Juelich Supercomputing Center. JUWELS Booster Supercomputer, 2020. https://apps.fz-juelich.de/jsc/hps/juwels/configuration.html#hardware-configuration-of-the-system-name-booster-module.

[19] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

[20] Apoorv Khandelwal, Luca Weihs, Roozbeh Mottaghi, and Aniruddha Kembhavi. Simple but effective: Clip embeddings for embodied ai. *arXiv preprint arXiv:2111.09888*, 2021.

[21] Alexander Kolesnikov, Lucas Beyer, Xiaohua Zhai, Joan Puigcerver, Jessica Yung, Sylvain Gelly, and Neil Houlsby. Big transfer (bit): General visual representation learning. In *European conference on computer vision*, pages 491–507. Springer, 2020.

[22] 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(1):32–73, 2017.

[23] Srivatsa Kundurthy. Lantern-rd: Enabling deep learning for mitigation of the invasive spotted lanternfly, 2022. URL https://arxiv.org/abs/2205.06397.

[24] Ang Li, Allan Jabri, Armand Joulin, and Laurens Van Der Maaten. Learning visual n-grams from web data. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 4183–4192, 2017.

[25] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation, 2022. URL https://arxiv.org/abs/2201.12086.

[26] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. *arXiv preprint arXiv:2201.12086*, 2022.

[27] 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 *European conference on computer vision*, pages 740–755. Springer, 2014.

[28] Dhruv Mahajan, Ross Girshick, Vignesh Ramanathan, Kaiming He, Manohar Paluri, Yixuan Li, Ashwin Bharambe, and Laurens Van Der Maaten. Exploring the limits of weakly supervised pretraining. In *Proceedings of the European conference on computer vision (ECCV)*, pages 181–196, 2018.

[29] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. Model cards for model reporting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT)*. ACM, 2019.

[30] Ron Mokady, Amir Hertz, and Amit H Bermano. Clipcap: Clip prefix for image captioning. *arXiv preprint arXiv:2111.09734*, 2021.

[31] Alex Nichol, Prafulla Dhariwal, Aditya Ramesh, Pranav Shyam, Pamela Mishkin, Bob McGrew, Ilya Sutskever, and Mark Chen. Glide: Towards photorealistic image generation and editing with text-guided diffusion models, 2021. URL `https://arxiv.org/abs/2112.10741`.

[32] Hieu Pham, Zihang Dai, Golnaz Ghiasi, Hanxiao Liu, Adams Wei Yu, Minh-Thang Luong, Mingxing Tan, and Quoc V Le. Combined scaling for zero-shot transfer learning. *arXiv preprint arXiv:2111.10050*, 2021.

[33] 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, 2021. URL `https://arxiv.org/abs/2103.00020`.

[34] 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 *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

[35] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. *CoRR*, abs/2102.12092, 2021. URL `https://arxiv.org/abs/2102.12092`.

[36] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents, 2022. URL `https://arxiv.org/abs/2204.06125`.

[37] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. *CoRR*, abs/2112.10752, 2021. URL `https://arxiv.org/abs/2112.10752`.

[38] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. doi: 10.1007/s11263-015-0816-y.

[39] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily Denton, Seyed Kamyar Seyed Ghasemipour, Burcu Karagol Ayan, S. Sara Mahdavi, Rapha Gontijo Lopes, Tim Salimans, Jonathan Ho, David J Fleet, and Mohammad Norouzi. Photorealistic text-to-image diffusion models with deep language understanding, 2022. URL `https://arxiv.org/abs/2205.11487`.

[40] Navid Kanaani Sajjad Ayoubi. Clipfa: Connecting farsi text and images. `https://github.com/SajjjadAyobi/CLIPfa`, 2021.

[41] Patrick Schramowski, Christopher Tauchmann, and Kristian Kersting. Can machines help us answering question 16 in datasheets, and in turn reflecting on inappropriate content? In *Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (FAccT)*. ACM, 2022.

[42] 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. *arXiv preprint arXiv:2111.02114*, 2021.

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

[44] Sheng Shen, Liunian Harold Li, Hao Tan, Mohit Bansal, Anna Rohrbach, Kai-Wei Chang, Zhewei Yao, and Kurt Keutzer. How much can clip benefit vision-and-language tasks? *arXiv preprint arXiv:2107.06383*, 2021.

[45] Makoto Shing. Japanese clip. https://github.com/rinnakk/japanese-clip, May 2022.

[46] Guijin Son, Hansol Park, Jake Tae, and Trent Oh. Koclip. https://github.com/jaketae/koclip, 20201.

[47] 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, 2021.

[48] Ryan Steed and Aylin Caliskan. Image representations learned with unsupervised pre-training contain human-like biases. In *Proceedings of ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, pages 701–713, 2021.

[49] Chen Sun, Abhinav Shrivastava, Saurabh Singh, and Abhinav Gupta. Revisiting unreasonable effectiveness of data in deep learning era. In *Proceedings of the IEEE international conference on computer vision*, pages 843–852, 2017.

[50] Rohan Taori, Achal Dave, Vaishaal Shankar, Nicholas Carlini, Benjamin Recht, and Ludwig Schmidt. Measuring robustness to natural distribution shifts in image classification. *Advances in Neural Information Processing Systems*, 33:18583–18599, 2020.

[51] Bart Thomee, David A Shamma, Gerald Friedland, Benjamin Elizalde, Karl Ni, Douglas Poland, Damian Borth, and Li-Jia Li. Yfcc100m: The new data in multimedia research. *Communications of the ACM*, 59(2):64–73, 2016.

[52] Maria Tsimpoukelli, Jacob L Menick, Serkan Cabi, SM Eslami, Oriol Vinyals, and Felix Hill. Multimodal few-shot learning with frozen language models. *Advances in Neural Information Processing Systems*, 34:200–212, 2021.

[53] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[54] Mitchell Wortsman, Gabriel Ilharco, Mike Li, Jong Wook Kim, Hannaneh Hajishirzi, Ali Farhadi, Hongseok Namkoong, and Ludwig Schmidt. Robust fine-tuning of zero-shot models. *arXiv preprint arXiv:2109.01903*, 2021.

[55] 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.

[56] Xiaohua Zhai, Joan Puigcerver, Alexander Kolesnikov, Pierre Ruyssen, Carlos Riquelme, Mario Lucic, Josip Djolonga, Andre Susano Pinto, Maxim Neumann, Alexey Dosovitskiy, et al. A large-scale study of representation learning with the visual task adaptation benchmark. *arXiv preprint arXiv:1910.04867*, 2019.

[57] Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers. 2021. doi: 10.48550/ARXIV.2106.04560. URL `https://arxiv.org/abs/2106.04560`.

[58] Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers. *arXiv preprint arXiv:2106.04560*, 2021.

[59] Xiaohua Zhai, Xiao Wang, Basil Mustafa, Andreas Steiner, Daniel Keysers, Alexander Kolesnikov, and Lucas Beyer. Lit: Zero-shot transfer with locked-image text tuning. *arXiv preprint arXiv:2111.07991*, 2021.

[60] Yinglin Zheng, Hao Yang, Ting Zhang, Jianmin Bao, Dongdong Chen, Yangyu Huang, Lu Yuan, Dong Chen, Ming Zeng, and Fang Wen. General facial representation learning in a visual-linguistic manner. *CoRR*, abs/2112.03109, 2021. URL `https://arxiv.org/abs/2112.03109`.