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

Christoph Schuhmann[1] §§°° **Romain Beaumont**[1] §§°° **Richard Vencu**[1,3,8] §§°°
Cade Gordon[2] §§°° **Ross Wightman**[1]§§ **Mehdi Cherti** [1,10]§§
Theo Coombes[1] **Aarush Katta**[1] **Clayton Mullis**[1] **Mitchell Wortsman**[6]
Patrick Schramowski[1,4,5] **Srivatsa Kundurthy**[1] **Katherine Crowson**[1,8,9]
Ludwig Schmidt[6] °° **Robert Kaczmarczyk**[1,7] °° **Jenia Jitsev**[1,10] °°
LAION[1] UC Berkeley[2] Gentec Data[3] TU Darmstadt[4] Hessian.AI[5]
University of Washington, Seattle[6] Technical University of Munich[7] Stability AI[8]
EleutherAI[9] Juelich Supercomputing Center (JSC), Research Center Juelich (FZJ)[10]
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 text-guided image generation and transfer to downstream tasks, while performing remarkably at zero-shot classification with noteworthy out-of-distribution robustness. Since then, large-scale language-vision models like ALIGN, BASIC, GLIDE, Flamingo and Imagen made further improvements. Studying the training and 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, GLIDE and Stable Diffusion 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 dataset exploration and subset generation, and detection scores for watermark, NSFW, and toxic content detection. [1]

1 Introduction

Learning from multimodal data such as text, images, and audio is a longstanding research challenge in machine learning [31, 51, 56, 83, 86]. Recently, contrastive loss functions combined with large neural networks have led to breakthroughs in the generalization capabilities of vision and language models [58, 59, 66]. For instance, OpenAI's CLIP models [58] achieved large gains in zero-shot classification on ImageNet [65], improving from the prior top-1 accuracy of 11.5% [41] to 76.2%.

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

In addition, CLIP achieved unprecedented performance gains on multiple challenging distribution shifts [3, 23, 61, 70, 78, 82]. Inspired by CLIP's performance, numerous groups have further improved image-text models by increasing the amount of computation and the training set size [28, 54, 89, 94]. Another recent success of multimodal learning is in image generation, where DALL-E [59] and later models [52, 60, 64, 66, 90] demonstrated the potential of text-guided image generation by producing high-quality images specific to the provided text.

A critical ingredient in this new generation of image-text models is the pre-training dataset. All of the aforementioned advances rely on large datasets containing hundreds of millions or even billions of image-text pairs, e.g., 400 million for CLIP [58] and 6.6 billion for BASIC [54]. However, *none of these datasets are publicly available*. While OpenAI still released the CLIP models publicly [58], later papers made neither the pre-training dataset nor the resulting models available to the wider research community [2, 28, 52, 54, 66, 89, 90]. As a result, research in this area has pooled into a small number of industrial research labs, limiting transparency and impeding research progress.

In this work, we address this challenge and make multimodal training more accessible by assembling a public dataset that is suitable for training large image-text models. Specifically, we introduce LAION-5B, the largest public image-text dataset containing over 5.8 billion examples (see Table 1 for a comparison). By starting from Common Crawl [1] and filtering this data source with an existing CLIP model, we derive a dataset consisting of three parts: 2.32 billion English image-text examples, 2.26 billion multilingual examples, and 1.27 billion examples that are not specific to a particular language (e.g., places, products, etc.). Beyond assembling the dataset, we also explore its ethical implications and flaws that emerge with large-scale data collection. By releasing LAION-5B publicly, we offer the first opportunity for the community to audit and refine a dataset of this magnitude.

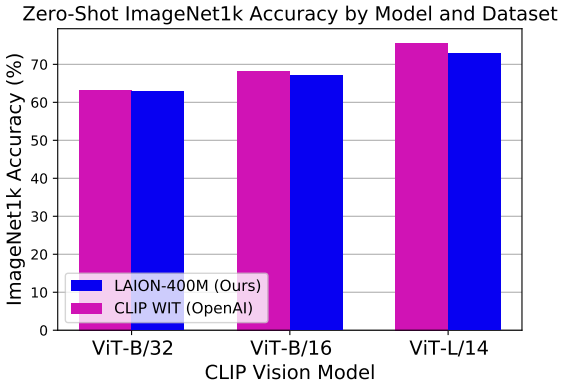


Figure 1: **Zero-Shot Accuracy.** CLIP models trained on LAION-400M (ours) [69], a previously released subset of LAION-5B, show competitive zero-shot accuracy compared to CLIP models trained on OpenAI's original training set WIT when evaluated on ImageNet1k.

Dataset	# English Img-Txt Pairs
Public Datasets	
MS-COCO	330K
CC3M	3M
Visual Genome	5.4M
WIT	5.5M
CC12M	12M
RedCaps	12M
YFCC100M	100M[2]
LAION-5B (Ours)	**2.3B**
Private Datasets	
CLIP WIT (OpenAI)	400M
ALIGN	1.8B
BASIC	6.6B

Table 1: **Dataset Size.** LAION-5B is more than 20 times larger than other public English image-text datasets. We extend the analysis from Desai et al. [14] and compare the sizes of public and private image-text datasets.

[2]Although YFCC100M contains 100M image-text pairs, it is unclear how well the text matches the image for an average example from the dataset. Radford et al. [57]'s curation procedure reduced YFCC100M to 15M samples.

To validate that LAION-5B is indeed suitable for training large image-text models, we conduct multiple experiments. We focus on matching the performance of OpenAI's CLIP models because they are the largest publicly released image-text models. OpenAI's CLIP models were trained on 400 million image-text pairs, and hence we also train CLIP models on a subset of LAION-5B containing the same number of examples ("LAION-400M"). Across a diverse range of problem settings including ImageNet (zero-shot), distribution shifts, VTAB, retrieval, and fine-tuning, our models trained on LAION-400M match or come close to the performance of OpenAI's CLIP models. Our ViT-L/14 models trained with OpenCLIP are the first open source reproductions of the largest CLIP models released by OpenAI.

Despite these validation results, 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. Hence we do not only release our dataset, but also our software stack we built for assembling LAION-5B. 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. We advise against any applications in deployed systems without carefully investigating behavior and possible biases of models trained on LAION-5B.**

The remainder of the paper proceeds as follows. After reviewing related work, we present our data collection process for LAION-5B in Section 3. Section 4 then describes LAION-5B's composition including its various subsets. To validate LAION-5B, we reproduce and evaluate different image-text models in Section 5. Before concluding, we discuss the technical limitations of LAION-5B in Section 6 and safety and ethics concerns in Section 7.

2 Related Work

Vision-Language Models. Radford et al. [58] made a large step forward in multimodal learning for image-text data with their CLIP (Contrastive Language–Image Pre-training) model. The authors proposed a contrastive learning scheme to embed both images and text into a shared representation space, which enabled unparalleled performance in zero-shot image classification. Moreover, CLIP made large progress on multiple challenging distribution shifts [78, 84].

After CLIP's initial success, ALIGN and BASIC improved contrastive multimodal learning by increasing the training set size and the batch size used for training [28, 54]. LiT also increased training scale and experimented with a combination of pre-trained image representations and contrastive fine-tuning to connect frozen image representations to text [94]. Flamingo introduced the first large vision-language model with in-context learning [2]. Other papers have combined contrastive losses with image captioning to further improve performance [43, 89]. Beyond image classification and retrieval, the community later adapted CLIP to further vision tasks such as object navigation and visual question answering [17, 32, 50, 72].

Another direction that has recently seen large progress in multimodal learning is text-guided image generation [47, 62, 95]. Specifically, DALL-E demonstrated diverse image generation capabilities for text prompts combining multiple concepts [59]. GLIDE, DALL-E 2, Imagen, Parti, and Stable Diffusion then improved visual fidelity and text-prompt correspondence [52, 60, 64, 66, 90].

Image-Text Datasets. Earlier dataset creation efforts such as MS-COCO and Visual Genome curated image and region labels through human annotation [36, 44]. While this resulted in high-quality labels, it also limited the scale of the datasets to only 330K and 5M examples, respectively. The web-harvested YFCC-100M dataset is substantially larger with about 99 million images and one million videos from Flickr, but only contains the user-generated metadata without additional annotations collected specifically for training computer vision models [79]. As a result, the text associated with an image sometimes has little to no correspondence with the actual image content.

To address this shortcoming of web-harvested image-text data, the Conceptual Captions dataset (CC3M) started with images and alt-text collected from the web, but then performed additional data cleaning procedures [71]. To increase the size of the dataset, researchers later relaxed the filtering protocol to arrive at the subsequent CC12M dataset [11]. Building datasets from alt-text continued with ALT200M [26] and ALIGN [28], which increased the dataset size up to 1.8 billion image-text pairs. In contrast to relying on alt-text, RedCaps used the captions provided by Reddit users to collect higher quality captions [14].

Datasets with non-English image-text pairs are less common. As a result, researchers translated English captioning datasets to other languages such as Farsi, Korean, and Japanese [67, 73, 74]. To the best of our knowledge, the largest multilingual dataset before LAION-5B has around 36 million samples from Wikipedia Image Text [75]. With the release of LAION-5B, researchers now have access to roughly two orders of magnitude more multilingual samples, which provides new opportunities for research on low-resource languages and multilingual models.

Scaling Behavior. Improving model performance by increasing data scale has been a theme in machine learning since at least the ImageNet dataset [13]. In the following decade, computer vision benefited from growth in model, data, and compute scale, in addition to advances in both convolutional and transformer architectures [15, 33, 81, 92]. Industrial research labs assembled large internal datasets such as Instagram-1B, JFT300M, and JFT3B to support image pre-training [46, 77, 93]. Natural language processing (NLP) demonstrated the beneficial effect of model, data, and compute scale on generalization through large language models such as GPT-3 [8] and associated experiments on scaling behavior [30]. Community efforts like the The Pile [18] and BigScience ROOTS [40] made large text datasets more accessible.

3 Collection Methodology

We constructed LAION-5B starting from Common Crawl, a public web archive [1]. The Common Crawl organization crawls the web since 2008 and publishes the results in snapshots approximately every month. Recent snapshots each contain about 300 TiB of data for around 3 billion web pages. In the following, we introduce our pipeline to assemble and filter a vision-language dataset from images in Common Crawl and their associated HTML alt-text.

3.1 Dataset Assembly Pipeline

Our dataset assembly pipeline follows the flowchart of Figure 2. At a high level, the pipeline consists of three main components: (i) distributed filtering of the Common Crawl web pages, (ii) distributed downloading of image-text pairs, and (iii) content filtering. The code used for the dataset pipeline

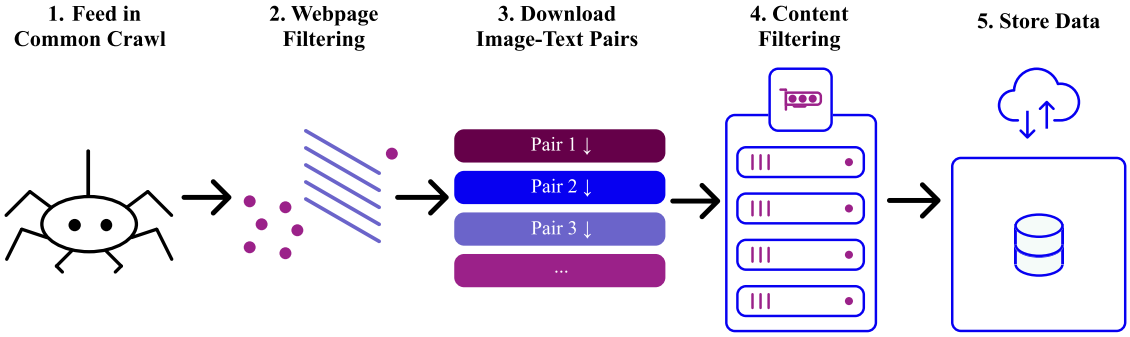


Figure 2: **Overview of the acquisition pipeline:** Files are downloaded, tracked, and undergo distributed inference to determine inclusion. Those above the specified CLIP threshold are saved.

may be found on GitHub[3]. We now describe each component in more detail.

Web page filtering. To extract image-text pairs from Common Crawl, we parse the HTML `IMG` (image) tags from Common Crawl's WAT metadata files.[4] Specifically, we focus on images with an *alt-text* so we can create image-text pairs. The alt-text is an HTML attribute of `IMG` tags containing alternative text for situations where the corresponding image cannot be rendered. For instance, screen reader software for a visually impaired person may read the alt-text in place of an image, or a search engine may use the alt-text to better index a web page without analyzing the actual image content.

After extracting the alt-text, we perform language detection using CLD3 [53] with three possible outputs: English, another language, or no detected language (i.e., all detections are below a confidence threshold [69]). Based on a manual inspection of a random sample, the "no language" set contains language-agnostic short form text such as the names of products and places.

We stored the resulting data in a PostgreSQL server for processing in the next stages of the pipeline. We maintained about 500M image URLs in the server at all times.

Downloading Image-Text Pairs. In order to maximize resource utilization, we downloaded the raw images from the parsed URLs with asynchronous requests using the Trio and Asks Python libraries. To limit costs, we chose a small cloud node with 2 vCPUs, 1GB of RAM, and 10Mbps download bandwidth as a worker instance. Such a worker can process 10,000 links in about $10 - 15$ minutes. We utilized roughly 300 workers in parallel and batched the workload into chunks of 10,000 links taken from the aforementioned PostgreSQL server.

Post-Processing. After downloading the WAT files from Common Crawl, we removed data with less than 5 characters of text, less than 5 KB of image data, and potentially malicious, large, or redundant images. To conclude the pipeline, we filtered image-text pairs based on their content. Specifically, we computed cosine similarities between the image and text encodings with OpenAI's ViT-B/32 CLIP model. For languages other than English, we utilized the multi-lingual CLIP ViT-B/32 from Carlsson et al. [10]. While OpenAI also released larger CLIP models later, these

[3] `https://github.com/rvencu/crawlingathome-gpu-hcloud`

[4] See `https://commoncrawl.org/the-data/get-started/` for details of the metadata format.

models were not available when we began to assemble LAION-5B. For consistency, we therefore relied on ViT-B/32 CLIP models for the entire dataset. We removed all English image-text pairs with cosine similarity below 0.28, and all other pairs with similarity below 0.26. This step removed around 90% of the original 50 billion images, leaving just short of 6 billion examples.

3.2 Safety During Collection

Current automated filtering techniques are far from perfect: harmful images are likely to pass, and others are likely to be falsely removed. We make a best effort to identify, document, and tag such content. In the case of illegal content, we computed CLIP embeddings to filter out such samples. Furthermore, these images and texts could amplify the social bias of machine learning models, especially ones trained with no or weak supervision [76]. 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. 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.

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 [49] which subsets and tagged images are used.

We apply Q16 [68] 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.* [68] 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.[5] 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.

4 Dataset Composition

We release LAION-5B as the following three subsets:

- 2.32 billion English image-text pairs. We refer to this subset as LAION-2B-en or LAION-2B if the language is clear from context.
- 2.26 billion image-text pairs from over 100 other languages. In the multilingual subset, the top-5 most frequent languages are Russian (10.6%), French (7.4%), German (6.6%), Spanish (6.6%), and Chinese (6.3%).

[5] 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).

- 1.27 billion samples where a language could not be clearly detected. Based on visually inspecting a random subset of these low-confidence language samples, the corresponding images often depict products or places. The captions contain language with clear semantics, but might also include noise such as keywords for search engine optimiziation or product tags.

We provide metadata files in the Apache Parquet format that consist of the following attributes for each image-text pair:

- A 64-bit integer identifier
- The URL of the image.
- The text string.
- Height and width of the image.
- Cosine similarity between the text and image embeddings.
- The output from our NSFW and watermark detectors (one score between 0 and 1 each).

3% of images were detected as NSFW, which can be filtered out by a user with the NSFW tag.

5 Experiments Validating LAION-5B

In this section, we showcase prior work using the LAION-400M [69] 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 enables novel dataset curation for 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 [96] and invasive species mitigation [38]. Within LAION, we have compiled from LAION-5B both LAION-High-Resolution[6], a 170M subset for superresolution models, and LAION-Aesthetic[7], a 120M subset of aesthetic images, as determined by a linear estimator on top of CLIP.

[6]https://huggingface.co/datasets/laion/laion-high-resolution
[7]https://github.com/LAION-AI/laion-datasets/blob/main/laion-aesthetic.md

CLIP Reproduction and Improvements. Gao et al. [19], trained an enhanced CLIP architecture on the LAION-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. Training on a LAION-5B subset, Li et al. [42] 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. [16] 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* [80].

Image Generation. Rombach et al. [63] utilized a subset of LAION-5B in training latent diffusion models (LDM) that achieved state-of-the-art results on image inpainting and class-conditional image synthesis. The work was further extended into stable diffusion project that used subsets of LAION-5B (LAION-2B-en, laion-high-resolution and laion-aesthetics[8]) for training a publicly available SOTA text-to-image generative model (see Appendix Sec. F.2). Furthermore, Gu et al. [21] used LAION-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. [66] showed an improved architecture of a diffusion model that was trained on a subset of LAION-400M that outperforms OpenAI's recent DALLE-2 and achieves 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 [58], and to validate the data collection pipeline we describe in Sec. 3, we trained several models on LAION-400M [69] 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, the usage of supercomputers and large compute clusters is necessary in order to train the models efficiently.

We used OpenCLIP [27], an open source software for training CLIP-like models. After adapting OpenCLIP for distributed training and execution on JUWELS Booster supercomputer [29], we reproduced CLIP models of different size on the LAION-400M subset. We trained ViT-B/32, ViT-B/16, and ViT-L/14 following CLIP [58], 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. All trained models may be found in the OpenCLIP repository[9]. 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 [58] 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 [58, 94]. 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.

[8] See `https://github.com/CompVis/stable-diffusion` for more details

[9] `https://github.com/mlfoundations/open_clip`

In Tab. 2, we show a comparison between models trained on LAION (400M, 2B) and original CLIP from [58]. We follow [94] and evaluate robustness performance on ImageNet distribution shift datasets [3, 23, 25, 61, 82]. Additionally, we construct a benchmark we call VTAB+, a superset of VTAB [91], on which we compute the average top-1 accuracy over 35 tasks[10]. 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 is better) and INet-S (LAION is better), which allows us to conclude that in overall, CLIP models trained on LAION match in their robustness original CLIP. With ViT-B/32 and ViT-L/14, training on the larger LAION-2B-en improves over LAION-400M model everywhere.

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-en). 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 [93]. 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.

To show that larger dataset scale matters for the performance of pre-trained models, we perform additional experiments using ViT-B/32 and ViT-L/14 on different LAION-5B and LAION-400M subsets, while varying the amount of training compute (samples seen). Our findings confirm that the effect of dataset scale is significant, given sufficient compute for training. For instance, for the same amount of compute (34B images seen), training ViT-L/14 on LAION-2B-en (75.4%) outperforms LAION-400M (73.9%) on ImageNet-1k zero-shot classification. Same effect is observed for smaller ViT-B/32 model. For more detailed results, see Fig. 12 and Tab. 6 in the Appendix.

5.3 Experiments with Generative Models

To validate LAION-5B as a dataset for training strong text-to-image generation models, we fine-tuned OpenAI's GLIDE [52] on LAION-5B data. The obtained results comparing generated samples from original OpenAI GLIDE and from our reproduction (LAIONIDE) are compiled into an interactive web demo[11]. See Appendix Sec F for more technical details on experiments with GLIDE (F.1) and Stable Diffusion (F.2).

6 Technical Limitations

The large scale of current image-text datasets makes it infeasible to thoroughly investigate all aspects of a dataset in a single publication. Hence we now outline some potential technical limitations specifically affecting LAION-5B. These potential limitations are starting points for future work on analyzing and improving image-text datasets.

[10] [91] 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.

[11] https://wandb.ai/afiaka87/glide_compare/reports/laionide-v3-benchmark–VmlldzoxNTg3MTkz

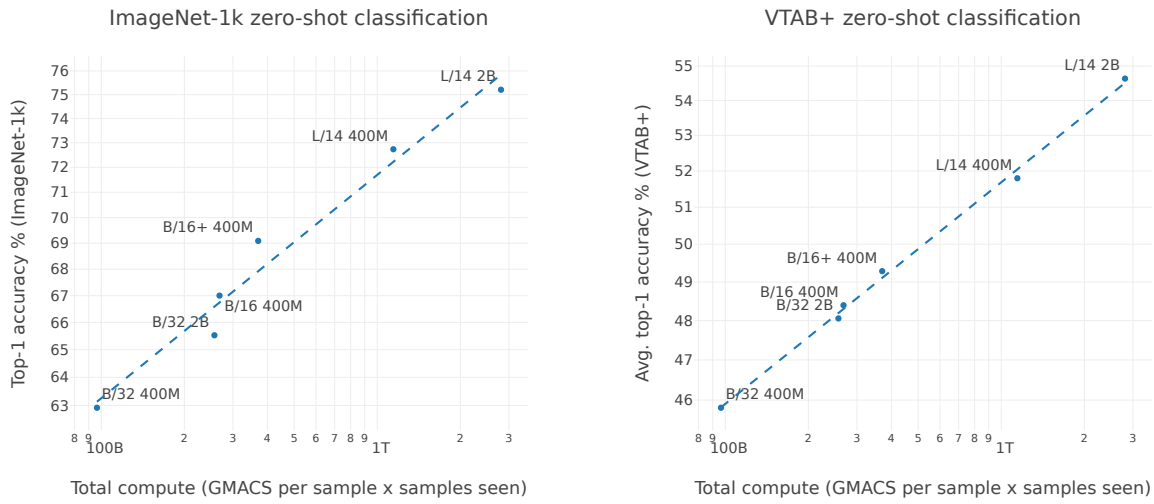


Figure 4: The relationship between total compute (giga multiply–accumulates (GMACS)) and zero-shot top-1 classification accuracy (%) of models trained on LAION (400M, 2B-en). The dashed line in each figure is a linear fit in log-log space. Each point corresponds to a model trained on either the 400M or 2B-en LAION subsets. We show results on ImageNet-1k (left) and VTAB+ (right) where we average the accuracy over 35 tasks (see Appendix E.3 for details). Clear effect of model, data and compute training scale is evident on zero-shot performance that increases following scale power law.

Data Overlap. Our experiments in Section 5.2 show that models trained on LAION-5B achieve good performance on a variety of downstream tasks. However, the LAION-5B training set may overlap with some of the downstream test sets if these test sets are also included in Common Crawl. If overlap is present, it may lead to incorrectly large test set accuracies that overstate the true generalization capabilities of models trained on LAION-5B.

Overall, we do not consider potential test set overlap to be a serious threat for the validity of results obtained with LAION-5B. OpenAI encountered the same question in the context of their pre-training dataset for CLIP and found only few examples of substantial performance difference due to data overlap on downstream target datasets [58]. Some datasets such as ObjectNet [3] are likely not contained in Common Crawl because ObjectNet was not assembled from web images. Instead, the authors of ObjectNet tasked MTurk workers to take new pictures in their own homes. Nevertheless, measuring the degree of overlap between LAION-5B and popular computer vision benchmarks is an important question for future work, which will include further de-duplication efforts.

Other text sources. Birhane et al. [6] described the shortcomings of alt-text and noted that alt-text is not necessarily a good description of the corresponding image. For instance, the alt-text may be search engine optimization (SEO) spam, an incoherent list of keywords, or overly corrupted otherwise. In such cases, the language in the text annotations may become less informative or entirely useless for training. For ImageNet zero-shot classification, BASIC [54] has demonstrated strong results when turning 5 billion of the 6.6 billion captions into the form of `CLASS_1 and CLASS_2 and ... and CLASS_K`, by using an internal multi-label classification dataset (JFT-3B). Thus, image captions

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$^{-3.9}$
	LAION-2B-en	75.2$^{-0.3}$	67.7$^{-2.0}$	87.4$^{-0.5}$	63.3$^{+3.7}$	65.5$^{-3.6}$	54.6$^{-1.2}$

Table 2: Comparison between CLIP models trained on LAION (400M, 2B) and the original CLIP models [58] 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.3 for more details about the datasets used for evaluation and the results.

formed by just concatenating class names may also serve as meaningful alternative of otherwise corrupted text. Such a finding adds a possibility of employing generated together with existing natural language captions for training contrastive image-language models with strong zero-shot performance.

Filtering with CLIP. CLIP allows the curation and collection of this dataset to be low-cost and scalable. Such an automated process reduces dramatically necessity for the human control which would be otherwise intractable for such large scale collection. However, through curating with CLIP, we also incur its flaws and model biases. For additional discussion of CLIP filtering related to safety and ethics, see Appendix Sec. G.2.

Filtering by a small scale CLIP ViT-B/32 may leave more image-text pairs with weak or no semantic connection in the dataset while also accidentally removing some high quality image-text pairs than filtering with stronger, larger scale models that were not available in the time of our experiments. The larger CLIP ViT-L/14 model may create a less noisy version of LAION datasets than what was possible with smaller scale CLIP ViT-B/32. We hypothesize that filtering Common Crawl with a CLIP ViT-L model will further increase the quality of our dataset. It is subject to our future work to create a CLIP ViT L/14 filtered version of LAION-400M and LAION-5B to test how this affects model training and downstream transfer performance.

7 Safety and Ethical Discussion

Recent developments in large-scale models, such as GPT-3 [9], CLIP [57], ALIGN [28], GLIDE [52], and DALLE-2 [60] have potential for far-reaching impact on society, both positive and negative, when deployed in applications such as image classification and generation, recommendation systems, or search engines. Besides model parameter scaling, the advances made so far also rely on the

underlying large-scale datasets. Recent research [4, 5] described many potential negative societal implications that may arise due to careless use of vision-language models, e.g., the models perform worse for certain groups of users or reproduce discriminatory behavior.

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. While open-source efforts exist to re-implement model architectures and training, the closed nature of large-scale datasets used for model training makes any proper systematic investigation of model training and model behavior very hard or even impossible. Studying full training, comparison of different model architectures and progress in large-scale multi-modal learning becomes restricted to those institutions that were able to obtain their closed large-scale datasets. It also results in safety issues of creating and using such models, as broad research community does not get to test both model and the dataset used for its training for causes underlying undesired behaviours.

LAION-5B as an open large-scale dataset provides here 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 under which circumstances their usage may result in undesired safety issues. Such research can help to design automated ways to curate and create datasets from uncurated ones that alleviate the bias and safety issues. To this end, LAION also created a number of tools to aid researchers and other users in large-scale data handling and exploration. One such a tool uses pre-computed image embeddings to enable search of images guided either by text or image input via an easily and publically accessible web interface (CLIP retrieval tool[12], see Appendix Sec. C.4). LAION made also source code for the tool and routines necessary to build an own version of it publicly available[13] (see Appendix Sec C, C.2, C.3 for more details).

After the release of LAION-400M, several groups (e.g., [6]) already used such tools and 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 safety and quality of the text-to-image models trained on the dataset.

Such development indicates that this dataset acts as a starting point, and is not the final endpoint, for creating further improved datasets to train models for various tasks. In our opinion, this process is not supposed to be a non-transparent closed-door avenue. It should be approached by broad research community, resulting in open and transparent datasets and procedures for model training. 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.

In the current form, we consider this dataset a research artefact and strongly advocate **academic use-only** and advise careful investigation of downstream model biases (Appendix Sec. G.2). Additionally, we encourage users to use the described tools and to transparently explore and, subsequently, report further not yet detected content and model behaviour to our dataset repository[14], and help to further advance existing approaches for data curation using the real-world large dataset introduced here.

[12] https://knn5.laion.ai
[13] https://github.com/rom1504/clip-retrieval
[14] https://github.com/laion-ai/laion5b-bias

Privacy. We comment on privacy issues arising from Common Crawl as source of links in LAION-5B and measures undertaken to handle those in the Appendix Sec. G.1

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.

Acknowledgments

We thank Phil Wang, the creator of the DALLE-pytorch github repository[15], 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, Gabriel Ilharco, Fredde Frallan (see also Appendix) and all the members of the LAION discord server[16] 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. We would also like to thank openAI for making their pre-trained CLIP models publicly available, which allowed us to filter the LAION datasets. We would like to express gratitude to all the people who are working on making code, models and data publicly available, advancing community based research and making research more reproducible.

The authors gratefully acknowledge the Gauss Centre for Supercomputing e.V. [17] for funding this work by providing computing time through the John von Neumann Institute for Computing (NIC) on the GCS Supercomputer JUWELS Booster [29] at Jülich Supercomputing Centre (JSC). We also acknowledge storage resources on JUST [20] granted and operated by JSC. Patrick Schramowski acknowledges the support by the Hessian Ministry of Higher Education, Research, Science and the Arts (HMWK) cluster project "The Third Wave of AI".

References

[1] URL `https://commoncrawl.org/`.

[15] `https://github.com/lucidrains/DALLE-pytorch`
[16] `https://discord.gg/xBPBXfcFHd`
[17] `https://gauss-centre.eu`

[2] 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.

[3] Andrei Barbu, David Mayo, Julian Alverio, William Luo, Christopher Wang, Dan Gutfreund, Josh Tenenbaum, and Boris Katz. Objectnet: A large-scale bias-controlled dataset for pushing the limits of object recognition models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. URL `https://proceedings.neurips.cc/paper/2019/file/97af07a14cacba681feacf3012730892-Paper.pdf`.

[4] 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.

[5] 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.

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

[7] Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. Food-101–mining discriminative components with random forests. In *European Conference on Computer Vision (ECCV)*, 2014. `https://data.vision.ee.ethz.ch/cvl/datasets_extra/food-101/`.

[8] 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.

[9] 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`.

[10] Fredrik Carlsson, Philipp Eisen, Faton Rekathati, and Magnus Sahlgren. Cross-lingual and multilingual clip. In *Proceedings of the Language Resources and Evaluation Conference*, pages 6848–6854, Marseille, France, June 2022. European Language Resources Association. URL `https://aclanthology.org/2022.lrec-1.739`.

[11] 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.

[12] Mircea Cimpoi, Subhransu Maji, Iasonas Kokkinos, Sammy Mohamed, and Andrea Vedaldi. Describing textures in the wild. In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2014. `https://arxiv.org/abs/1311.3618`.

[13] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pages 248–255. Ieee, 2009.

[14] 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.

[15] 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.

[16] 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`.

[17] Samir Yitzhak Gadre, Mitchell Wortsman, Gabriel Ilharco, Ludwig Schmidt, and Shuran Song. CLIP on wheels: Zero-shot object navigation as object localization and exploration, 2022.

[18] 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.

[19] 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`.

[20] 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.

[21] 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`.

[22] Varun Gulshan, Lily Peng, Marc Coram, Martin C. Stumpe, Derek Wu, Arunachalam Narayanaswamy, Subhashini Venugopalan, Kasumi Widner, Tom Madams, Jorge Cuadros, Ramasamy Kim, Rajiv Raman, Philip C. Nelson, Jessica L. Mega, and Dale R. Webster. Development and Validation of a Deep Learning Algorithm for Detection of Diabetic Retinopathy in Retinal Fundus Photographs. *JAMA*, 316(22):2402–2410, 12 2016. ISSN 0098-7484. doi: 10.1001/jama.2016.17216. URL `https://doi.org/10.1001/jama.2016.17216`.

[23] Dan Hendrycks, Steven Basart, Norman Mu, Saurav Kadavath, Frank Wang, Evan Dorundo, Rahul Desai, Tyler Zhu, Samyak Parajuli, Mike Guo, Dawn Song, Jacob Steinhardt, and Justin Gilmer. The many faces of robustness: A critical analysis of out-of-distribution generalization. *International Conference on Computer Vision (ICCV)*, 2021. `https://arxiv.org/abs/2006.16241`.

[24] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 15262–15271, 2021.

[25] Dan Hendrycks, Kevin Zhao, Steven Basart, Jacob Steinhardt, and Dawn Song. Natural adversarial examples. *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2021. https://arxiv.org/abs/1907.07174.

[26] 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.

[27] 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.

[28] 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.

[29] 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.

[30] 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.

[31] 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.

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

[33] 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.

[34] Simon Kornblith, Jonathon Shlens, and Quoc V Le. Do better imagenet models transfer better? In *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2019. https://arxiv.org/abs/1805.08974.

[35] Jonathan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *International Conference on Computer Vision (ICCV) Workshops*, 2013. https://ieeexplore.ieee.org/document/6755945.

[36] 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.

[37] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images, 2009. https://www.cs.toronto.edu/~kriz/learning-features-2009-TR.pdf.

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

[39] Alina Kuznetsova, Hassan Rom, Neil Alldrin, Jasper Uijlings, Ivan Krasin, Jordi Pont-Tuset, Shahab Kamali, Stefan Popov, Matteo Malloci, Alexander Kolesnikov, Tom Duerig, and Vittorio Ferrari. The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale. *IJCV*, 2020.

[40] 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, et al. The bigscience roots corpus: A 1.6 tb composite multilingual dataset. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

[41] 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.

[42] 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.

[43] 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.

[44] 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.

[45] Luping Liu, Yi Ren, Zhijie Lin, and Zhou Zhao. Pseudo numerical methods for diffusion models on manifolds. *arXiv preprint arXiv:2202.09778*, 2022.

[46] 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.

[47] Elman Mansimov, Emilio Parisotto, Jimmy Lei Ba, and Ruslan Salakhutdinov. Generating images from captions with attention. *arXiv preprint arXiv:1511.02793*, 2015.

[48] Maxim Maximov, Ismail Elezi, and Laura Leal-Taixé. Ciagan: Conditional identity anonymization generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 5447–5456, 2020.

[49] 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.

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

[51] Yasuhide Mori, Hironobu Takahashi, and Ryuichi Oka. Image-to-word transformation based on dividing and vector quantizing images with words. In *First international workshop on multimedia intelligent storage and retrieval management*, pages 1–9. Citeseer, 1999.

[52] 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`.

[53] Jeroen Ooms. *cld3: Google's Compact Language Detector 3*, 2022. https://docs.ropensci.org/cld3/, https://github.com/ropensci/cld3 (devel) https://github.com/google/cld3 (upstream).

[54] 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.

[55] Jordi Pont-Tuset, Jasper Uijlings, Soravit Changpinyo, Radu Soricut, and Vittorio Ferrari. Connecting vision and language with localized narratives. In *ECCV*, 2020.

[56] Ariadna Quattoni, Michael Collins, and Trevor Darrell. Learning visual representations using images with captions. In *2007 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1–8. IEEE, 2007.

[57] 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`.

[58] 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.

[59] 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`.

[60] 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`.

[61] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do ImageNet classifiers generalize to ImageNet? In *International Conference on Machine Learning (ICML)*, 2019. `https://arxiv.org/abs/1902.10811`.

[62] Scott Reed, Zeynep Akata, Xinchen Yan, Lajanugen Logeswaran, Bernt Schiele, and Honglak Lee. Generative adversarial text to image synthesis. In *International conference on machine learning*, pages 1060–1069. PMLR, 2016.

[63] 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.

[64] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021.

[65] 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.

[66] 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.

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

[68] 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.

[69] 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.

[70] Vaishaal Shankar, Achal Dave, Rebecca Roelofs, Deva Ramanan, Benjamin Recht, and Ludwig Schmidt. Do image classifiers generalize across time?, 2019. https://arxiv.org/abs/1906.02168.

[71] 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.

[72] 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.

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

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

[75] 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.

[76] 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.

[77] 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.

[78] 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.

[79] 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.

[80] 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.

[81] 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.

[82] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019. `https://arxiv.org/abs/1905.13549`.

[83] Jason Weston, Samy Bengio, and Nicolas Usunier. Large scale image annotation: learning to rank with joint word-image embeddings. *Machine learning*, 81(1):21–35, 2010.

[84] 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.

[85] Jianxiong Xiao, Krista A Ehinger, James Hays, Antonio Torralba, and Aude Oliva. Sun database: Exploring a large collection of scene categories. *International Journal of Computer Vision*, 2016. `https://link.springer.com/article/10.1007/s11263-014-0748-y`.

[86] Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *International conference on machine learning*, pages 2048–2057. PMLR, 2015.

[87] Kaiyu Yang, Jacqueline H Yau, Li Fei-Fei, Jia Deng, and Olga Russakovsky. A study of face obfuscation in imagenet. In *International Conference on Machine Learning*, pages 25313–25330. PMLR, 2022.

[88] 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`.

[89] 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.

[90] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2022.

[91] 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.

[92] 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`.

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

[94] 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.

[95] Han Zhang, Tao Xu, Hongsheng Li, Shaoting Zhang, Xiaogang Wang, Xiaolei Huang, and Dimitris N Metaxas. Stackgan: Text to photo-realistic image synthesis with stacked generative adversarial networks. In *Proceedings of the IEEE international conference on computer vision*, pages 5907–5915, 2017.

[96] 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`.

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

A Datasheet for LAION-5B dataset

A.1 Motivation

Q1 **For what purpose was the dataset created?** Was there a specific task in mind? Was there a specific gap that needed to be filled? Please provide a description.

- LAION-5B was created as an open solution to training very large multimodal models such as CLIP or DALL-E. Before the curation of this dataset, the closest in size was YFCC with 100 million image/videos and associated metadata. OpenAI previously used a 15 million sample subset to train a publicly comparable CLIP model, but that pales in comparison to the private 400 million sample dataset they used to train the high-performant CLIP models. At the time of writing this, the ImageNet-1k zero-shot top-1 state-of-the-art, Google's BASIC, used a dataset of 6.6 billion image-text pairs. With the release of LAION-5B, researchers no longer have to be part of a few selected institutions to study these problems.

Q2 **Who created the dataset (e.g., which team, research group) and on behalf of which entity (e.g., company, institution, organization)?**

- This dataset is presented by LAION (Large-scale Artificial Intelligence Open Network), a non-profit research organization aiming to democratize access to large-scale open datasets and powerful machine learning models through the research and development of open-source resources. The communication and organization of this project took place on the open LAION discord server [18].

Q3 **Who funded the creation of the dataset?** If there is an associated grant, please provide the name of the grantor and the grant name and number.

- This work was sponsored by Hugging Face and Stability AI.

Q4 **Any other comments?**

- No.

A.2 Composition

Q5 **What do the instances that comprise the dataset represent (e.g., documents, photos, people, countries)?** *Are there multiple types of instances (e.g., movies, users, and ratings; people and interactions between them; nodes and edges)? Please provide a description.*

- We provide 5.8 billion image-text pairs. Each pair consists of the following: an image file url; text caption; width; height; the caption's language; cosine similarity (CLIP ViT/B-32 for English and MCLIP for multiple and unknown languages); the probability of the image containing a watermark; the probability of a sample being NSFW. We made our models

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

openly available on the LAION github page (`https://github.com/LAION-AI/LAION-5B-WatermarkDetection`, `https://github.com/LAION-AI/CLIP-based-NSFW-Detector`).

Q6 **How many instances are there in total (of each type, if appropriate)?**

- LAION-5B contains 2.3 billion English samples, 2.2 billion multilingual samples, and 1.2 billion unknown language samples. A further overview of the statistics may be seen in the announcement blog post .

Q7 **Does the dataset contain all possible instances or is it a sample (not necessarily random) of instances from a larger set?** *If the dataset is a sample, then what is the larger set? Is the sample representative of the larger set (e.g., geographic coverage)? If so, please describe how this representativeness was validated/verified. If it is not representative of the larger set, please describe why not (e.g., to cover a more diverse range of instances, because instances were withheld or unavailable).*

- Common Crawl is a public repository of crawled web pages. From this collection of web pages we filter the images and alt-text to derive LAION-5B. Of the existing 50+ billion images available in common crawl. We provide image url and alt-text pairings of only 5.8 billion images.

Q8 **What data does each instance consist of?** *"Raw" data (e.g., unprocessed text or images) or features? In either case, please provide a description.*

- We provide raw urls and their associated alt-text.

Q9 **Is there a label or target associated with each instance?** *If so, please provide a description.*

- There is no hard class label, but researchers will often formulate a mapping of the text to image or vice-versa.

Q10 **Is any information missing from individual instances?** *If so, please provide a description, explaining why this information is missing (e.g., because it was unavailable). This does not include intentionally removed information, but might include, e.g., redacted text.*

- No.

Q11 **Are relationships between individual instances made explicit (e.g., users' movie ratings, social network links)?** *If so, please describe how these relationships are made explicit.*

- No.

Q12 **Are there recommended data splits (e.g., training, development/validation, testing)?** *If so, please provide a description of these splits, explaining the rationale behind them.*

- No.

Q13 **Are there any errors, sources of noise, or redundancies in the dataset?** *If so, please provide a description.*

- There exist near duplicate images which makes possible a many to one embedding in certain scenarios. CLIP embeddings may be used to remove more or less of them.

Q14 **Is the dataset self-contained, or does it link to or otherwise rely on external resources (e.g., websites, tweets, other datasets)?** *If it links to or relies on external resources, a) are there guarantees that they will exist, and remain constant, over time; b) are there official archival versions of the complete dataset (i.e., including the external resources as they existed at the time the dataset was created); c) are there any restrictions (e.g., licenses, fees) associated with any of the external resources that might apply to a future user? Please provide descriptions of all external resources and any restrictions associated with them, as well as links or other access points, as appropriate.*

- This dataset is reliant on links to the World Wide Web. As such, we are unable to offer any guarantees of the existence of these samples. Due to the size we will also not be able to offer archives of the current state either. In order to rapidly and efficiently download images from URLs, we provide img2dataset. Depending on bandwidth, it's feasible to download the entire LAION-5B dataset in 7 days using 10 nodes.

Q15 **Does the dataset contain data that might be considered confidential (e.g., data that is protected by legal privilege or by doctor–patient confidentiality, data that includes the content of individuals' non-public communications)?** *If so, please provide a description.*

- This dataset was collected using openly available parts of the internet with the assumption that any data found was intended to be shared freely. However, it is possible that the parties crawled by Common Crawl may have publicly hosted confidential data.

Q16 **Does the dataset contain data that, if viewed directly, might be offensive, insulting, threatening, or might otherwise cause anxiety?** *If so, please describe why.*

- Since the dataset is scraped from Common Crawl, it is known to have instances of sexually explicit, racist, abusive or other discomforting or disturbing content. We choose to include these samples for the usage of safety researchers and further dataset curation surrounding these sensitive topics.

- To address the existence of distressing content, we provide safety tags. Details on tagging potentially inappropriate content can be found in Sec. 3.2 in the main text and Appendix Sec. C.5 and Sec. C.6. During down-stream training tasks, users may check the sample's boolean flags to determine whether or not the sample should be used. However, as we described in the main text, it is important to note that the safety tags are not perfect, especially keeping the complexity of these tasks and the diverse opinions of different cultures in mind. 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.

Q17 **Does the dataset relate to people?** *If not, you may skip the remaining questions in this section.*

- People may be present in the images or textual descriptions, but people are not the sole focus of the dataset.

Q18 **Does the dataset identify any subpopulations (e.g., by age, gender)?**

- We do not provide any markers of subpopulation as attributes of the image-text pairs, but it may be possible to deduce this in some cases from the image and language pairing.

Q19 **Is it possible to identify individuals (i.e., one or more natural persons), either directly or indirectly (i.e., in combination with other data) from the dataset?** *If so, please describe how.*

- Yes it may be possible to identify people using face recognition. We do not provide any such means nor make attempts, but institutions owning large amounts of face identifiers may identify specific people in the dataset. Similarly, people may be identified through the associated text.

Q20 **Does the dataset contain data that might be considered sensitive in any way (e.g., data that reveals racial or ethnic origins, sexual orientations, religious beliefs, political opinions or union memberships, or locations; financial or health data; biometric or genetic data; forms of government identification, such as social security numbers; criminal history)?** *If so, please provide a description.*

- Yes the dataset contains sensitive content. Although, the dataset wasn't created with the intention of obtaining samples fitting this criteria, it is possible that individuals might have hosted such items on a website that had been crawled by Common Crawl.

Q21 **Any other comments?**

- We caution discretion on behalf of the user and call for responsible usage of the dataset for research purposes **only**.

A.3 Collection Process

Q22 **How was the data associated with each instance acquired?** *Was the data directly observable (e.g., raw text, movie ratings), reported by subjects (e.g., survey responses), or indirectly inferred/derived from other data (e.g., part-of-speech tags, model-based guesses for age or language)? If data was reported by subjects or indirectly inferred/derived from other data, was the data validated/verified? If so, please describe how.*

- From the aforementioned Common Crawl, we filter images and their associated alt-text. Inclusion is determined by cosine similarity of the alt-text and the image as determined by OpenAI's CLIP ViT-B/32 for english samples and MCLIP for all other samples. We include English samples with a cosine similarity score above 0.28, and we select all multilingual and unknown language samples with a 0.26 cosine similarity score or greater.

Q23 **What mechanisms or procedures were used to collect the data (e.g., hardware apparatus or sensor, manual human curation, software program, software API)?** *How were these mechanisms or procedures validated?*

- We ran a preprocessing script in python, over hundred of small CPU nodes, and few GPU nodes. They were validated by manual inspection of the results and post processing on them: computation of statistics on the width, height, size of captions, clip embeddings and indices

Q24 **If the dataset is a sample from a larger set, what was the sampling strategy (e.g., deterministic, probabilistic with specific sampling probabilities)?**

- The dataset was obtained by openAI CLIP ViT B/32 filtering of Common Crawl links using cosine similarity of the image and its text the links were referring to.

Q25 **Who was involved in the data collection process (e.g., students, crowdworkers, contractors) and how were they compensated (e.g., how much were crowdworkers paid)?**

- No crowdworkers were used in the curation of the dataset. Open-source researchers and developers enabled its creation for no payment.

Q26 **Over what timeframe was the data collected? Does this timeframe match the creation timeframe of the data associated with the instances (e.g., recent crawl of old news articles)?** *If not, please describe the timeframe in which the data associated with the instances was created.*

- The data was filtered from September 2021 to January 2022, but those who created the sites might have included content from before then. It is impossible to know for certain how far back the data stretches.

Q27 **Were any ethical review processes conducted (e.g., by an institutional review board)?** *If so, please provide a description of these review processes, including the outcomes, as well as a link or other access point to any supporting documentation.*

- We corresponded with the University of Washington's Human Subject Division, and as we do not intervene with the people depicted in the data as well as the data being public, they stated that the work did not require IRB review. Furthermore, the NeurIPS ethics review determined that the work has no serious ethical issues.

Q28 **Does the dataset relate to people?** *If not, you may skip the remaining questions in this section.*

- People may appear in the images and descriptions, although they are not the exclusive focus of the dataset.

Q29 **Did you collect the data from the individuals in question directly, or obtain it via third parties or other sources (e.g., websites)?**

- We retrieve the data from Common Crawl which contains almost all websites.

Q30 **Were the individuals in question notified about the data collection?** *If so, please describe (or show with screenshots or other information) how notice was provided, and provide a link or other access point to, or otherwise reproduce, the exact language of the notification itself.*

- Individuals were not notified about the data collection.

Q31 **Did the individuals in question consent to the collection and use of their data?** *If so, please describe (or show with screenshots or other information) how consent was requested*

and provided, and provide a link or other access point to, or otherwise reproduce, the exact language to which the individuals consented.

- We follow Common Crawl's practice of crawling the web and follow each site's robots.txt file, thus users consent to their sites being crawled. However, those depicted in the photograph might not have given their consent to its upload.

Q32 **If consent was obtained, were the consenting individuals provided with a mechanism to revoke their consent in the future or for certain uses?** *If so, please provide a description, as well as a link or other access point to the mechanism (if appropriate).*

- Users have a possibility to check for the presence of the links in our dataset leading to their data on public internet by using the search tool provided by LAION, accessible at https://knn5.laion.ai. If users wish to revoke their consent after finding sensitive data, they can contact the hosting party and request to delete the content from the underlying website – it will be automatically removed from LAION-5B since we distributed image-text pairs as URLs. Moreover, we provide a contact email `contact@laion.ai` and contact form https://laion.ai/dataset-requests/ to request removal of the links from the dataset. The actual content behind the links is out of our reach and will in that case remain accessible on the public internet for other crawlers.

Q33 **Has an analysis of the potential impact of the dataset and its use on data subjects (e.g., a data protection impact analysis) been conducted?** *If so, please provide a description of this analysis, including the outcomes, as well as a link or other access point to any supporting documentation.*

- Birhane, Prabhu, and Kahembwe opened the discussion on the limitations and imminent biases that come with the creation of a weakly-curated dataset using CLIP. CLIP and its usage of cosine similarity offers a useful but imperfect heuristic for dataset inclusion that inherits various biases contained in the image-text pairs crawled from the web. In addition, the biases already existent within CLIP and the World Wide Web may become amplified when distilling original raw data and forming a filtered dataset. Using a model trained on this dataset without any further curation in production has the potential to reinforce harmful simplistic stereotypes against already marginalized communities.
- However, the authors also note that this dataset posits currently the only openly available solution for studying multimodal models of this scale, examining their potential benefits and harms. Combining the aforementioned limitations and opportunities that this dataset provides, we agree with the authors and authorize the dataset for purely academic endeavors and strongly advice against any usage in end products.

Q34 **Any other comments?**

- No.

A.4 Preprocessing, Cleaning, and/or Labeling

Q35 **Was any preprocessing/cleaning/labeling of the data done (e.g., discretization or bucketing, tokenization, part-of-speech tagging, SIFT feature extraction, removal**

of instances, processing of missing values)? *If so, please provide a description. If not, you may skip the remainder of the questions in this section.*

- No preprocessing or labelling is done. Certain images were removed on the basis of safety, and others are tagged in the presence of NSFW content or a watermark.

Q36 **Was the "raw" data saved in addition to the preprocessed/cleaned/labeled data (e.g., to support unanticipated future uses)?** *If so, please provide a link or other access point to the "raw" data.*

- We do not save the raw data.

Q37 **Is the software used to preprocess/clean/label the instances available?** *If so, please provide a link or other access point.*

- To preprocess the data we used:
  - `https://github.com/rvencu/crawlingathome-gpu-hcloud` process common crawl into a laion5B-like dataset
  - `http://github.com/rom1504/img2dataset` A tool to easily turn large sets of image urls to an image dataset. Can download, resize and package 100M urls in 20h on one machine.
  - `https://github.com/rom1504/clip-retrieval` a tool to easily compute clip embeddings and build a clip retrieval system with them
- For individuals to preprocess the data for training, we provide:
  - `https://github.com/rom1504/laion-prepro`

Q38 **Any other comments?**

- No.

A.5 Uses

Q39 **Has the dataset been used for any tasks already?** *If so, please provide a description.*

- LAION-5B (and the associated LAION-400M) has been used on a number of tasks such as CLIP Reproduction, BLIP Training, Glide Training, Cloob Training, and sub-dataset generation. For example, Gu et al. used LAION-400M to train VQ diffusion text-to-image generation models. Additionally, Rombach et al. applied a subset of LAION-400M in training Latent Diffusion Models that achieved state-of-the-art results on image inpainting and class-conditional image synthesis. The team behind open_CLIP demonstrated the capabilities of the 400M subset for CLIP reproduction, achieving performance on par with that of OpenAI. On the matter of subset generation and CLIP reproduction, Zheng et al. utilized LAION for facial representation learning. It should be noted that this example demonstrates the potential for users to misuse this dataset for the purpose of identification. Li et al. applied a subset of LAION for the purpose of image-captioning. Finally, Eichenberg et al. used a LAION subset for MAGMA, a model generating text "answers" for image-question pairs.

Q40 **Is there a repository that links to any or all papers or systems that use the dataset?** *If so, please provide a link or other access point.*

- Yes, scientific publications and systems that use LAION datasets can be found on the LAION github page.

Q41 **What (other) tasks could the dataset be used for?**

- We encourage future researchers to curate LAION-5B for several tasks. Particularly, we see applications of the dataset in image and text representation learning, image to text generation, image captioning, and other common multimodal tasks. Due to the breadth of the data, it also offers a unique opportunity for safety and low resource language researchers. We hope for LAION-5B to serve under-represented projects as well.

Q42 **Is there anything about the composition of the dataset or the way it was collected and preprocessed/cleaned/labeled that might impact future uses?** *For example, is there anything that a future user might need to know to avoid uses that could result in unfair treatment of individuals or groups (e.g., stereotyping, quality of service issues) or other undesirable harms (e.g., financial harms, legal risks) If so, please provide a description. Is there anything a future user could do to mitigate these undesirable harms?*

- As this data stems from the greater internet, it mirrors the broader biases of society in the period of its collection. Biases in subpopulation depiction (eg. correlation between gender and jobs), violence, and nudity (for which we provide safety tags) might create harmful outcomes for those a model might be applied to. For this reason this dataset should not be used to make a decision surrounding people.

Q43 **Are there tasks for which the dataset should not be used?** *If so, please provide a description.*

- Due to the known biases of the dataset, under no circumstance should any models be put into production using the dataset as is. It is neither safe nor responsible. As it stands, the dataset should be solely used for research purposes in its uncurated state.
- Likewise, this dataset should not be used to aid in military or surveillance tasks.

Q44 **Any other comments?**

- No.

A.6 Distribution

Q45 **Will the dataset be distributed to third parties outside of the entity (e.g., company, institution, organization) on behalf of which the dataset was created?** *If so, please provide a description.*

- Yes, the dataset will be open-source.

Q46 **How will the dataset be distributed (e.g., tarball on website, API, GitHub)?** *Does the dataset have a digital object identifier (DOI)?*

- The data will be available through Huggingface datasets.

Q47 **When will the dataset be distributed?**

- 31/03/2022 and onward.

Q48 **Will the dataset be distributed under a copyright or other intellectual property (IP) license, and/or under applicable terms of use (ToU)?** *If so, please describe this license and/or ToU, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms or ToU, as well as any fees associated with these restrictions.*

- CC-BY-4.0

Q49 **Have any third parties imposed IP-based or other restrictions on the data associated with the instances?** *If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any relevant licensing terms, as well as any fees associated with these restrictions.*

- LAION owns the metadata and release as CC-BY-4.0.
- We do not own the copyright of the images or text.

Q50 **Do any export controls or other regulatory restrictions apply to the dataset or to individual instances?** *If so, please describe these restrictions, and provide a link or other access point to, or otherwise reproduce, any supporting documentation.*

- No.

Q51 **Any other comments?**

- No.

A.7 Maintenance

Q52 **Who will be supporting/hosting/maintaining the dataset?**

- Huggingface will support hosting of the metadata.
- The Eye supports hosting of the embeddings and backups of the rest.
- LAION will maintain the samples distributed.

Q53 **How can the owner/curator/manager of the dataset be contacted (e.g., email address)?**

- `https://laion.ai/dataset-requests/`

Q54 **Is there an erratum?** *If so, please provide a link or other access point.*

- There is no erratum for our initial release. Errata will be documented as future releases on the dataset website.

Q55 **Will the dataset be updated (e.g., to correct labeling errors, add new instances, delete instances)?** *If so, please describe how often, by whom, and how updates will be communicated to users (e.g., mailing list, GitHub)?*

- LAION-5B will not be updated. However a future LAION-streamed-from-CC may exist for updates. Specific samples can be removed on request.

Q56 **If the dataset relates to people, are there applicable limits on the retention of the data associated with the instances (e.g., were individuals in question told that their data would be retained for a fixed period of time and then deleted)?** *If so, please describe these limits and explain how they will be enforced.*

- People may contact us at the LAION website to add specific samples to a blacklist.

Q57 **Will older versions of the dataset continue to be supported/hosted/maintained?** *If so, please describe how. If not, please describe how its obsolescence will be communicated to users.*

- We will continue to support LAION-400M.

Q58 **If others want to extend/augment/build on/contribute to the dataset, is there a mechanism for them to do so?** *If so, please provide a description. Will these contributions be validated/verified? If so, please describe how. If not, why not? Is there a process for communicating/distributing these contributions to other users? If so, please provide a description.*

- Unless there are grounds for significant alteration to certain indexes, extension of the dataset will be carried out on an individual basis.

Q59 **Any other comments?**

- No.

B Dataset Setup Procedure

After processing and filtering common crawl, 5B of image url/text samples are available. Here we provide an overview of all the steps necessary to combine the full dataset.

1. Downloading the data as webdataset with distributed img2dataset
2. Computing Vit-L/14 embeddings with distributed clip-inference
3. Computing a KNN index from these embeddings using autofaiss
4. Computing additional tags (NSFW and watermark) using CLIP embeddings

C Dataset Preparation and Curation Details

C.1 Distributed img2dataset

We developed img2dataset library to easily download, resize, and store images and captions in the webdataset format.[19] This allows to download 100 million images from our list of URLs in 20 hours with a single node (1Gbps connection speed, 32GB of RAM, an i7 CPU with 16 cores), allowing anyone to obtain the whole dataset or a smaller subset.

[19] https://github.com/rom1504/img2dataset

For LAION-5B we introduced a distributed mode for this tool, allowing to download the 5B samples in a week using 10 nodes. see [20] and [21]

C.2 Distributed CLIP inference

From these images, the CLIP retrieval inference tool [22] was used to compute ViT-L/14 embeddings, allowing for a better analysis capacity of the data. In particular a distributed mode [23] made it possible to compute these embeddings in a week using 32 NVIDIA A100s: this larger CLIP model can only be computed at a speed of 312 sample/s per gpu, compared to 1800 sample/s for ViT-B/32.

The resulting embeddings are available for everyone to use for clustering, indexing, linear inference.

C.3 Distributed indexing

We then used these 9TB of image embeddings to build a large PQ128 knn index using the autofaiss tool [24]. To make this run faster, a distributed mode is available [25]

C.4 Integration in the search UI

In order to demonstrate the value of this data, we integrated this index into the [26] UI. It is powered by the code called clip back at [27] The knn index is 800GB and the metadata (url and captions) as well, so memory mapping is used for both in order to use no RAM, only a SSD drive of that capacity is required.

C.5 Specialized NSFW image content tagging

We applied various tagging to the content of LAION 5B. Among other contents, we tagged images with pornographic or sexualized content (referred to as NSFW). To ensure all implementations related to LAION-5B are open-source, we refrained from using existing commercial solutions.

In particular, we first trained an EfficientNetV2-based classifier. However, then moved to a simple MLP based on OpenAI's CLIP/L-14. To this end, we created a training dataset by retrieving images from the previous LAION-400M dataset which are close in the CLIP embedding space to various keywords related to the five categories: "neutral", "drawing", "porn", "hentai" or "sexy". Additionally, we added SFW images from the Wikiart [28] and Danbooru datasets [29] to the "drawing" category and NSFW images from Danbooru to the "hentai" category.

Following this procedure, we obtained over 682K images from the five classes "drawing" (39026), "hentai" (28134), "neutral" (369507), "porn" (207969) and "sexy" (37914). Using this data we trained

[20] https://github.com/rom1504/img2dataset/blob/main/dataset_examples/laion5B.md
[21] https://github.com/rom1504/img2dataset/blob/main/examples/distributed_img2dataset_tutorial.md
[22] https://github.com/rom1504/clip-retrieval
[23] https://github.com/rom1504/clip-retrieval/blob/main/docs/distributed_clip_inference.md
[24] https://github.com/criteo/autofaiss
[25] https://github.com/criteo/autofaiss/blob/master/docs/distributed/distributed_autofaiss.md
[26] https://knn5.laion.ai
[27] https://github.com/rom1504/clip-retrieval
[28] https://www.wikiart.org
[29] https://www.gwern.net/Danbooru2021

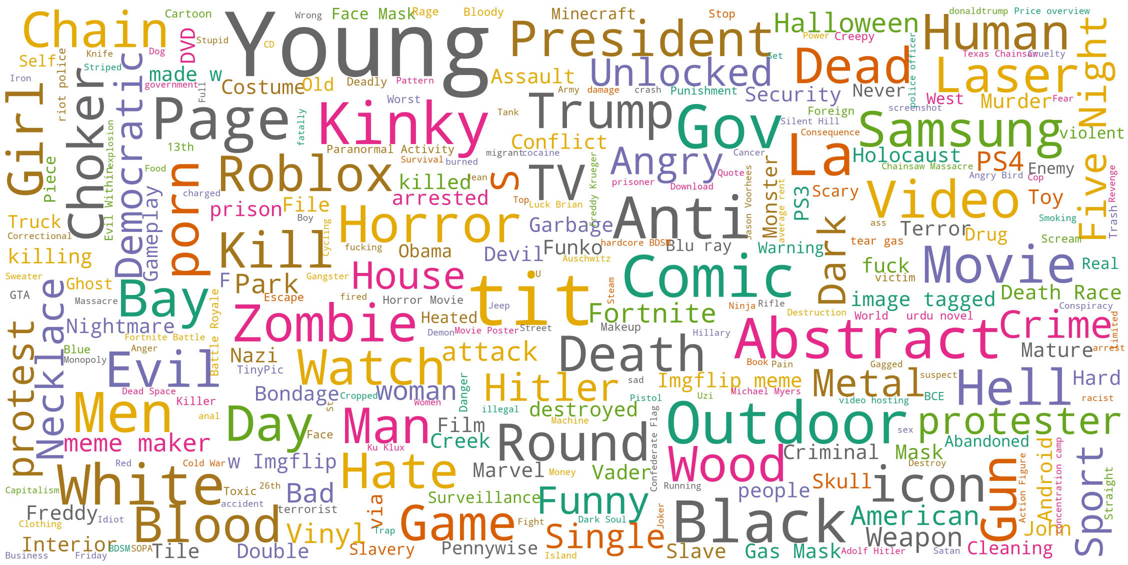


Figure 5: Word cloud based on [68] documenting the potentially inappropriate image content of the LAION-5B subset which contains text in English language. Provided alternative text is used as text description of the images. Word size is proportional to the word counts and rank in descriptions corresponding to the inappropriate image set.

a detector for these five categories by finetuning an ImageNet-1k pretrained EfficientNet-V2-B02 model. [30] To use this image classifier as a binary SFW - NSFW classifier, we consider images from the classes "drawing" and "neutral" as SFW and "hentai", "porn" and "sexy" as NSFW. To measure the performance of this model, we created a test dataset with 1000 images from each category and manually inspected it, to make sure all test images where correctly annotated. Our EfficientNet-V2-B02 image classifier predicted 96,45% of the true NSFW correctly as NSFW and discards 7,96% of the SFW images incorrectly as NSFW.

C.6 Further inappropriate content tagging

Further, we used the Q16 documentation pipeline [68] to document the broad range of identified potentially inappropriate concepts contained, cf. Sec. 3.2 for details. Fig. 5 shows the most frequent identified concepts following this procedure. One can see that in a lot of cases these images show humans (cf. concepts *human, people, man, woman*). Further, one main concept is pornographic content (e.g. *porn, bondage, kinky, bdsm*). Additionally, most frequent present concepts are, among other concepts, *weapons, violence, terror, murder, slavery, racism* and *hate*. Note that also content surrounding *halloween* (*costume, halloween, zombie*) and art or media such as *movies*, *games* and *comics* are potentially tagged, depending on the displayed content. Further filtering depends highly on the use-case and users' opinions.

[30]Code may be found at: https://github.com/LAION-AI/LAION-SAFETY

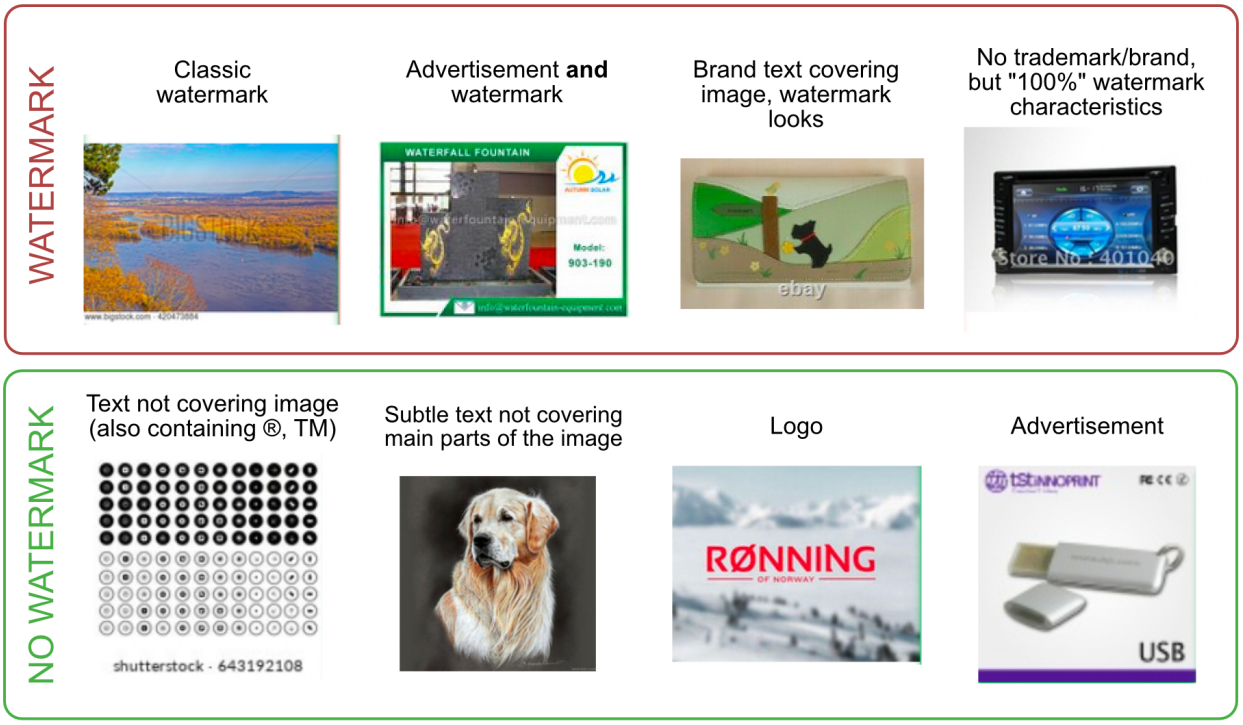


Figure 6: **Watermark test set annotation examples.** Criteria for LAION-5B sample annotation for watermark (top row) and non-watermark (bottom row) images.

C.7 Watermark and safety inference

Finally, we wanted to let user the ability to remove unsafe examples, and watermarked examples. To do that we collected training and test sets. The training set was augmented with examples retrieved from the KNN index, while the test set samples were selected to represent well the dataset distribution but were all manually annotated. 6

The inference is done using the embedding-reader[31] module.

These tags were then integrated in the UI, allowing everyone to observe that the safety tags indeed filter out almost all the unsafe results, and giving confidence that training a generative model on this data will not result in unexpectedly unsafe images.

D Dataset Samples and Statistics

Here, we present samples from the dataset and some distribution statistics to aid in understanding the dataset. In Figure 7, we randomly select 4 samples from each of the 3 LAION-5B subsets. As can be seen, the language classifier seems to have low confidence with names, identifying numbers, and short form text. An important future line of work will be to improve the language classifier.

To comprehend the dataset beyond visual examples, we may look at statistics collected about the

[31] https://github.com/rom1504/embedding-reader

English

Blue Beach Umbrellas, Point Of Rocks, Crescent Beach, Siesta Key - Spiral Notebook

BMW-M2-M-Performance-Dekor-Long-Beach-Blue-05

Becoming More Than a Good Bible Study Girl: Living the Faith after Bible Class Is Over [...]

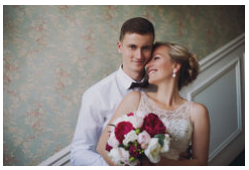
"Dynabrade 52632 4-1/2"" Dia. Right Angle Depressed Center Wheel Grinder (Replaces 50306 and 50346)"

Multilingual

Peugeot 308 2013 sedan

Episcopia Ortodoxa a Maramuresului si Satmarului are un nou Arhiereu vicar

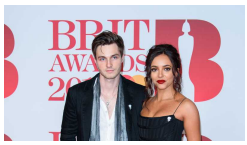
DON QUIJOTE DE LA MANCHA (SELECCIÓN DE TEXTOS)



Żeński i męski portret Dama outdoors i facet Ślubna para [...]

Low Confidence Language

18fcd9e025205 Fila Omnispeed Men Us 10 Multi Color Running Shoe in Blue for Men - Lyst



Little Mix's Jade Thirlwall has 'split' from her boyfriend Jed Elliot

Saarinen Style: M70 Womb Ottoman Mcm Furniture, Selling Furniture, [...]

Europe, Italy

Figure 7: **LAION-5B random examples from all subsets.** We take the first 4 SFW samples from each of the 3 randomly shuffled LAION-5B subsets. We present the image and its associated caption.

distribution. Figure 8 gives an overview of the caption length amongst all subsets. Additionally, Figure 9 describes the frequency of languages within the multilingual subset. The 10 most frequent languages compose 56% of the multilingual dataset.

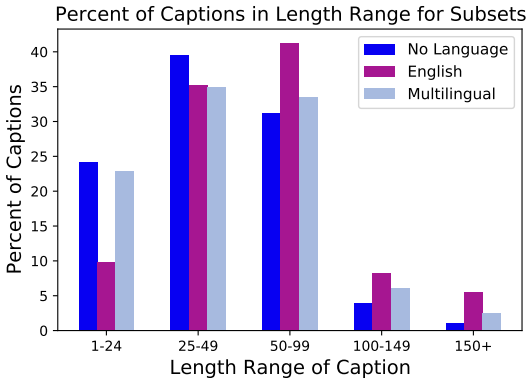


Figure 8: **Caption Character Length.** Each of the LAION-5B subsets contains similar frequencies and exhibit a right skew.

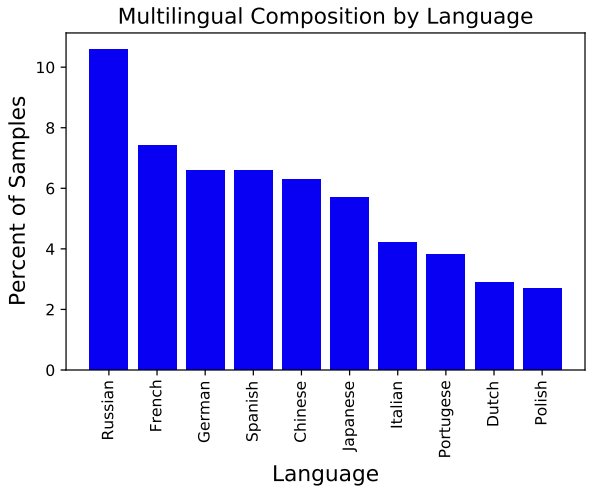


Figure 9: **Multilingual Language Frequency.** The 10 most frequent languages seem to be largely of European and East Asian origin.

E Further Experimental Details and Results on CLIP reproduction

We provide details about experiments that were done to reproduce CLIP [58] using LAION (400M, 2B-en) subsets. In addition, we document all experimental results on both zero-shot classification using the VTAB+ suite and retrieval.

E.1 Training Details

We used distributed data parallel training (using PyTorch DDP) to train models on multiple NVIDIA A100 GPUs. Training was done using the InfoNCE loss like in [58]. We used Adam with decoupled weight regularization (i.e., AdamW) as an optimizer, with $\beta_1 = 0.9$ and $\beta_2 = 0.98$ for all models. We used a linear warmup followed by a cosine decay schedule. For regularization we used the same weight decay of 0.2 for all the models. Details about different architectures that were used are provided in Tab. 3. Training hyper-parameters and resources used are provided in Tab. 4.

E.2 Distributed Training and InfoNCE Loss

To properly deal with global batch for contrastive InfoNCE loss in distributed setting, we need additional communication between GPU workers to compute the loss and the gradients for all positive and negative sample pairs correctly. In each worker, we gather all image and text embeddings from the other workers, and use them as negative examples for each image-text pair in the mini-batch.

A naive implementation of InfoNCE involves materializing a very large $N \times N$ matrix, N being the global batch size. For $N = 32768$, the matrix occupies a hefty 8 GB in float32. To remedy this, we use a formulation of the loss like OpenAI [58] where redundant operations are sharded to local devices while maintaining correct global gradients. This successfully overcomes a significant scaling issue and achieves a memory complexity that scales linearly with global batch size by only materializing 2 matrices of size $n \times N$, n being local batch size per GPU. By turning memory

complexity from $\mathcal{O}(N^2)$ into $\mathcal{O}(nN)$, we slash memory overhead due to scaling from GBs down to MBs.

Name	Width	Embed Dim	Depth	Res.	Acts.	Params
ViT-B/32	768 / 512	512	12 / 12	224x224	10 M	151 M
ViT-B/16	768 / 512	512	12 / 12	224x224	29 M	150 M
ViT-B/16+	896 / 640	640	12 / 12	240x240	40 M	208 M
ViT-L/14	1024 / 768	768	24 / 12	224x224	97 M	428 M

Table 3: Hyper-parameters of different architectures we used for reproducing CLIP models. **Acts** refers to the number of activations in millions, while **Params** refers to the number of parameters in millions. All entries in the form of A / B denote image and text parameters respectively.

Model (data size)	BS. (global)	#GPUs	LR.	Warm.	Ep.	Time (hrs.)
B/32 (400M)	256 (32768)	128	5e-4	2K	32	36
B/32 (2B)	416 (46592)	112	5.5e-4	10K	16	210
B/16 (400M)	192 (33792)	176	5e-4	2K	32	61
B/16+(400M)	160 (35840)	224	7e-4	5K	32	61
L/14 (400M)	96 (38400)	400	6e-4	5K	32	88

Table 4: Training hyper-parameters and resources used to reproduce CLIP [58] models on LAION 400M and 2B subsets. Note that **BS** refer to batch size per GPU worker (with **global** the corresponding global batch size), **LR** to base learning rate, **Warm** to the total number of warmup steps, **Ep** to the total number of training epochs, and **Time** to total training time in hours.

E.3 Detailed Results & Further Analysis

In this section we present all zero-shot classification results on VTAB+ as well as retrieval results. In Tab. 5, we describe the datasets that are used in VTAB+. For zero-shot classification, we collected prompts and class names from prior works [58, 94] and made them available in our benchmark repository[32]. In Tab. 7, we show zero-shot top-1 classification accuracy (%) on VTAB+ datasets. Tables 8 and 9 depict retrieval results on Flickr30K[88] and MSCOCO [44].

Effect of data scale. We observe similar or better results on most datasets when using the larger LAION-2B-en instead of LAION-400M. Exceptions are on some datasets with specialized domains (e.g., Diabetic Retinopathy, PatchCamelyon) or in structured tasks (see corresponding paragraph below). To demonstrate the importance of the data scale for the quality of the pre-trained models, we conduct a series of experiments where we vary both data scale (LAION-80M, LAION-400M and LAION-2B) and amount of training compute measured in samples seen (3B, 13B and 34B). We observe that when investing enough into training compute, seeing same number of samples on larger data scale leads consistently to better zero-shot transfer performance measured on ImageNet-1k. This is valid for both smaller B/32 and larger L/14 model scales. For instance, models pre-trained on LAION-2B outperform there significantly models pre-trained on LAION-400M, when using same

[32]https://github.com/LAION-AI/CLIP_benchmark

Dataset	Abbr.(Tab. 2, 7)	Test size	#Classes
ImageNet-1k	INet	50,000	1,000
ImageNet-v2	INet-v2	10,000	1,000
ImageNet-R	INet-R	30,000	200
ImageNet Sketch	INet-S	50,889	1,000
ObjectNet	ObjNet	18,574	113
ImageNet-A	INet-A	7,500	200
CIFAR-10	-	10,000	10
CIFAR-100	-	10,000	100
MNIST	-	10,000	10
Oxford Flowers 102	Flowers102	6,149	102
Stanford Cars	Cars	8,041	196
SVHN	-	26,032	10
Facial Emotion Recognition 2013	FER2013	7,178	7
RenderedSST2	-	1,821	2
Oxford-IIIT Pets	Pets	3,669	37
Caltech-101	-	6,085	102
Pascal VOC 2007 Classification	VOC2007-Cl	14,976	20
SUN397	-	108,754	397
FGVC Aircraft	-	3,333	100
Country211	-	21,100	211
Describable Textures	DTD	1,880	47
GTSRB	-	12,630	43
STL10	-	8,000	10
Diabetic Retinopathy	Retino	42,670	5
EuroSAT	-	5,400	10
RESISC45	-	6,300	45
PatchCamelyon	PCAM	32,768	2
CLEVR Counts	-	15,000	8
CLEVR Object Distance	CLEVR Dist	15,000	6
DSPRITES Orientation	DSPRITES Orient	73,728	40
DSPRITES Position	DSPRITES pos	73,728	32
SmallNORB Elevation	SmallNORB Elv	12,150	9
SmallNORB Azimuth	SmallNORB Azim	12,150	18
DMLAB	-	22,735	6
KITTI closest vehicle distance	KITTI Dist	711	4

Table 5: Datasets used for zero-shot classification evaluation (VTAB+).

large compute training budget of 34B samples seen (see Fig. 12 and Tab. 6). We conclude from these findings that extending dataset scale all the way up towards LAION-2B is indeed important for obtaining stronger zero-shot transfer performance, given sufficiently large compute for training.

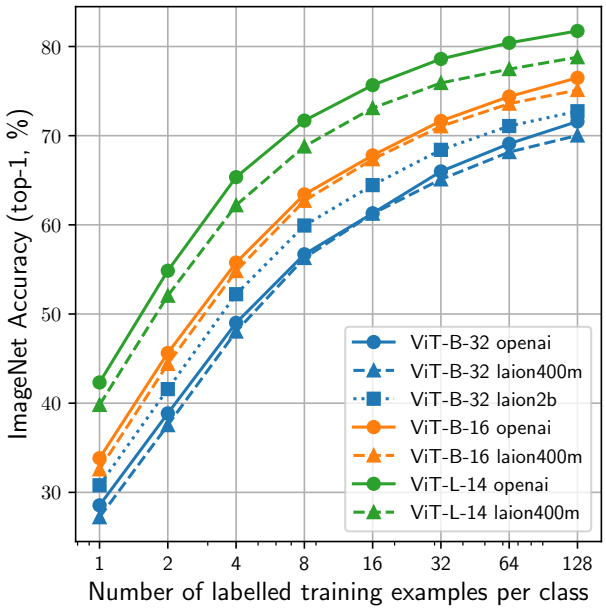


Figure 10: Evaluating few-shot linear probe performance on ImageNet. We evaluate i) models trained on various LAION subsets and ii) the original CLIP models. Models trained on LAION show similar transfer performance to those trained by OpenAI. Also evident is clear effect of model or data scale on transfer across few-shot conditions.

Few-shot transfer: comparison to CLIP and effect of scale. To examine the quality of the learned representations, we evaluate few-shot linear probe performance on seven datasets commonly used to benchmark transfer performance. The results are presented in Figures 10 and 11. Figure 10 displays few-shot performance on ImageNet [13] while Figure 11 displays few-shot performance on Food101 [7], Cars [35], CIFAR-10 & 100 [37], DTD [12] and SUN397 [85]. In addition to evaluating models trained on subsets of LAION, we also compare with the CLIP models of Radford *et al.* [58]. Overall we observe that the models trained on LAION achieve similar transfer performance to those trained by OpenAI. Moreover, we observe that performance increases with more data (i.e., B/32 2B outperforms B/32 400M) and larger models.

ImageNet-A In ImageNet-A [24] (noted INet-A), we observe large differences between CLIP WIT and LAION models, e.g. a difference of 24.3% on ViT-L/14. We note that INet-A design and data collection is quite different from other ImageNet distribution shifts datasets, as the images were specifically selected to be adversarial for a ResNet-50 pre-trained on ImageNet-1k. Although we do not have yet an explanation for the observed discrepancies and it would be interesting to understand why LAION models are worse than CLIP WIT, it is not clear whether improvements in INet-A are generalizable, as the dataset is based on adversarial images specific to a pre-trained model (ResNet-50).

Diabetic Retinopathy We observe a large variation of performance on Diabetic Retinopathy [22] (noted Retino). Accuracy goes from 3% to 73.3% for CLIP WIT models, and from 7.4% to 24.2% for LAION models. Additionally, the difference between CLIP WIT and LAION models goes up to

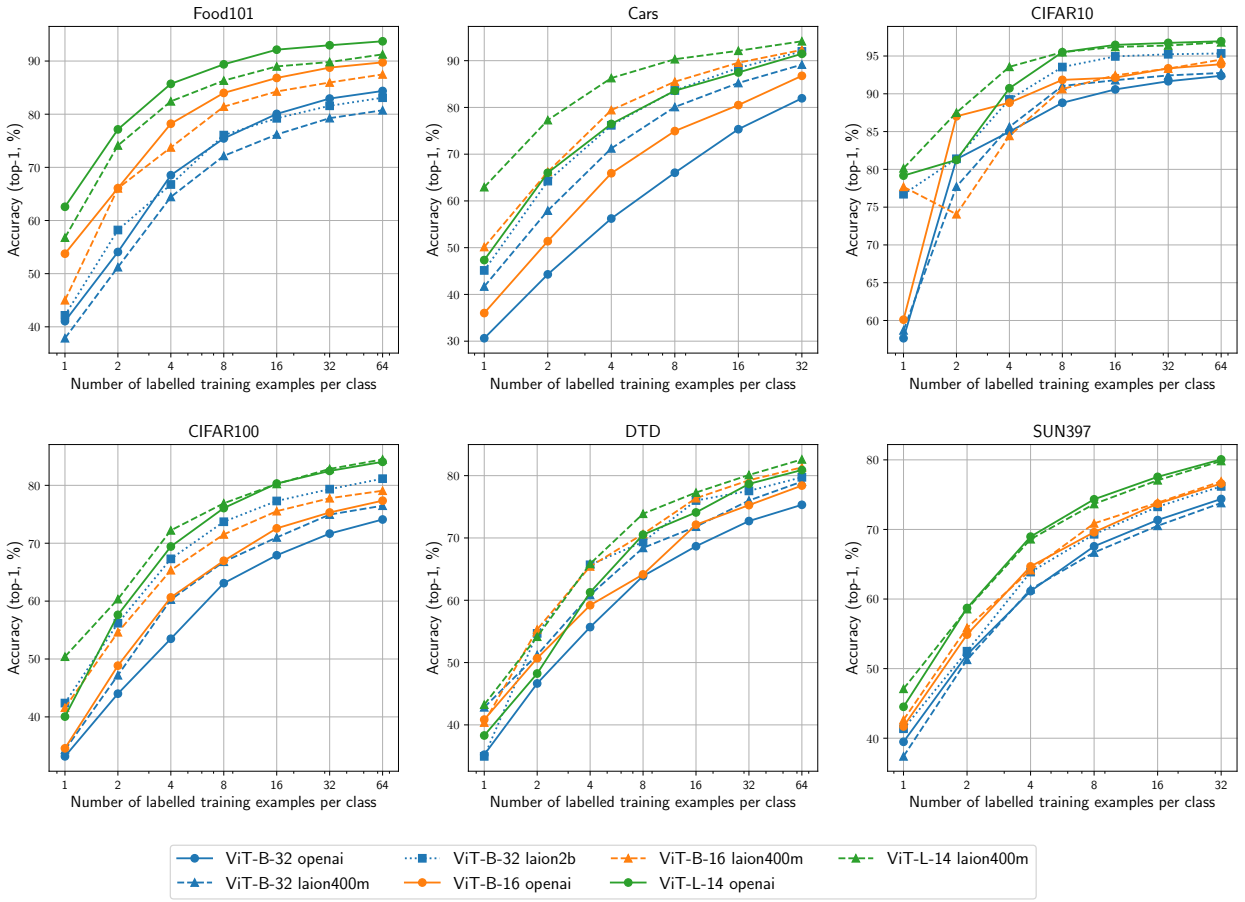


Figure 11: Evaluating few-shot linear probe performance on 6 datasets commonly used to benchmark transfer [34]. We evaluate i) models trained on various LAION subsets and ii) the original CLIP models. We evaluate performance on Food101 [7], Cars [35], CIFAR-10 & 100 [37], DTD [12] and SUN397 [85].

67.3% (on L/14). After investigating, we found that on low accuracy models, performance on the majority class is very low (e.g., for ViT-B/16 LAION model, recall was 3.4% on the majority class), and given that the dataset is highly imbalanced (majority class constitutes 74% of the samples), accuracy is affected heavily. A possible reason for low performance could be the prompts that were used, thus tuning the prompts could alleviate the problem. We re-evaluated the models using mean per-class recall, and found that the performances are less disparate, with a maximum difference between CLIP WIT models and LAION models of 2.1%. Overall, the results remain quite low, best mean per-class recall was 25.4%, obtained with ViT-B/32 trained on LAION-400M.

Structured tasks Similarly to [94], we observe low accuracy on VTAB's structured tasks [91] (CLEVR, DSPRITES, SmallNORB, DMLAB, KITTI) which involve counting, depth prediction, or position/angle prediction. Finding ways to improve accuracy on those tasks is an open research question [94] that would be interesting to investigate in future work.

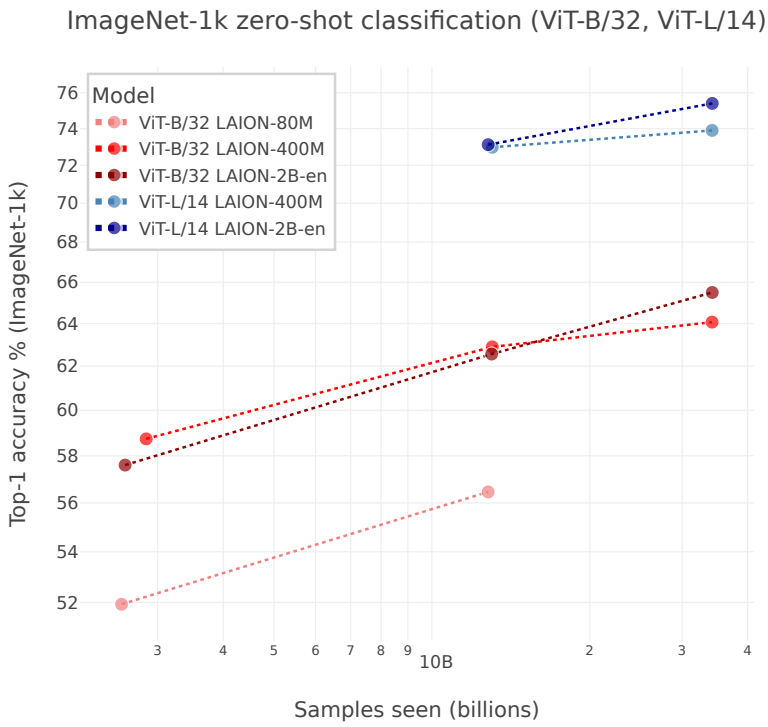


Figure 12: ViT-B/32 and ViT-L/14 additional experiments where we vary the amount compute (3B, 13B, and 34B images seen) and LAION subset size (80M, 400M, 2B). We evaluate the models on zero-shot Imagenet-1k classification. Seeing same number of samples on larger data scale leads consistently to better zero-shot transfer performance, when investing enough into training compute.

Retrieval We observe consistent improvements of LAION models over CLIP WIT models on MSCOCO 5K test set (Tab. 9) across all metrics and model sizes. On Flickr30k (Tab 8), we observe similar or better results with LAION models, with the exception of image retrieval on ViT-B/16 where CLIP WIT model is better. It would be interesting to investigate why LAION models have an advantage, and whether the advantage is more general or specific to the datasets that are considered in this work. Overall, we obtain better results than the best reported results in [58], e.g. on MSCOCO text retrieval we obtain 59.3% vs 58.4% for CLIP WIT, and on image retrieval we obtain 42% vs 37.8% for CLIP WIT, both evaluated using the R@1 metric.

F Overview of Experiments and Results on Generative Models

Here we provide overview about training experiments that were performed with generative models, GLIDE and Stable Diffusion, using subsets of LAION-5B.

F.1 GLIDE

OpenAI released checkpoints for the GLIDE [52] architecture to the public, but only released checkpoints trained on a filtered dataset removing hate-symbols and humans. These models can do a lot, but are incapable of generating imagery of humans. To evaluate the LAION dataset and its

Model	Samples seen	LAION-80M	LAION-400M	LAION-2B-en
ViT-B/32	3B	51.93	58.73	57.60
	13B	56.46	62.90	62.56
	34B	-	64.07	65.50
ViT-L/14	13B	-	72.98	73.12
	34B	-	73.90	75.40

Table 6: ViT-B/32 and ViT-L/14 additional experiments where we vary the amount compute (3B, 13B, and 34B images seen) and LAION subset size (80M, 400M, 2B). We evaluate the models on zero-shot Imagenet-1k classification. When investing enough into training compute, seeing same number of samples on larger data scale leads consistently to better zero-shot transfer performance measured on ImageNet-1k.

generalization capabilities, we aim to re-introduce the ability to generate imagery of humans into these checkpoints by finetuning them on LAION-5B.

We finetune the released GLIDE 64 pixel base (filtered) checkpoint from OpenAI on LAION-5B. For upscaling from 64x64 images to 256x256 images, we use the unmodified weights from OpenAI GLIDE-upsample-filtered. During training, captions were randomly replaced with the unconditional token 20% of the time. All code and checkpoints are provided in our GitHub repository [33].

We finetune LAIONIDE-v1 first, using an NVIDIA RTX 2070 Super GPU. Due to the 8GB VRAM constrain posed by the RTX 2070, we only use a batch size of 1. This initial checkpoint is provided as LAIONIDE-v1.

To accelerate training, LAIONIDE-v2 is finetuned from LAIONIDE-v1 using an 8xA100 pod from Stability. LAIONIDE-v2 sees roughly 25 million shuffled text-image pairs from LAION-2B. Some data is filtered during finetuning: if a text-image pair's 'nsfw' metadata has a value of 'NSFW' or 'LIKELY', we remove the sample. We remove any pairs where the language code is not 'en', to focus the model on english. We remove any images with an aspect ratio greater than 1.3 or less than 0.8. We remove all images where the smallest side is less than 256 pixels in length. Finally, we perform a sub-string search against a list of common slurs, and remove captions containing some slurs, although this is far from comprehensive.

To reduce the number of watermarks output by LAIONIDE-v2, we finetune to create LAIONIDE-v3. It sees roughly 1 million text-image pairs from a shuffled mixture of datasets: COCO 2017's training set (MS-COCO), Visual Genome, Open Images "Localized Annotations" and LAION-5B [36, 39, 44, 55]. We find this reduces the number of watermarks output compared to LAIONIDE-V2 during manual analysis.

To improve inference time we make use of the pseudo linear multi-step diffusion sampling method from Liu et al. [45] as implemented by Katherine Crowson.

We compare some evaluations from OpenAI's released filtered checkpoint and the one we train. Those can be found at the following link: `https://wandb.ai/afiaka87/glide_compare/reports/laionide-v3-benchmark--VmlldzoxNTg3MTkz`

[33] `https://github.com/LAION-AI/laionide`

Dataset	B/32			B/16		B/16+	L/14	
	CLIP WIT	LAION-400M	LAION-2B	CLIP WIT	LAION-400M	LAION-400M	CLIP WIT	LAION-400M
INet	63.3	62.9$^{-0.4}$	65.7$^{+2.4}$	68.3	67.0$^{-1.3}$	69.2	75.6	72.8$^{-2.8}$
INet-v2	56.0	55.1$^{-0.9}$	57.4$^{+1.4}$	61.9	59.6$^{-2.3}$	61.5	69.8	65.4$^{-4.4}$
INet-R	69.4	73.4$^{+4.0}$	75.9$^{+6.5}$	77.7	77.9$^{+0.2}$	80.5	87.9	84.7$^{-3.2}$
INet-S	42.3	49.4$^{+7.1}$	52.9$^{+10.6}$	48.2	52.4$^{+4.2}$	54.4	59.6	59.6
ObjNet	44.2	43.9$^{-0.3}$	48.7$^{+4.5}$	55.3	51.5$^{-3.8}$	53.9	69.0	59.9$^{-9.1}$
INet-A	31.6	21.7$^{-9.9}$	26.1$^{-5.5}$	49.9	33.2$^{-16.7}$	36.9	70.8	46.5$^{-24.3}$
CIFAR-10	89.8	90.7$^{+0.9}$	94.0$^{+4.2}$	90.8	91.7$^{+0.9}$	92.7	95.6	94.6$^{-1.0}$
CIFAR-100	64.2	70.3$^{+6.1}$	75.4$^{+11.2}$	66.9	71.2$^{+4.3}$	73.8	75.9	77.4$^{+1.5}$
MNIST	48.2	37.4$^{-10.8}$	63.4$^{+15.2}$	51.8	66.3$^{+14.5}$	57.0	76.4	76.0$^{-0.4}$
Flowers102	66.5	68.1$^{+1.6}$	69.0$^{+2.5}$	71.2	69.3$^{-1.9}$	71.1	79.2	75.6$^{-3.6}$
Cars	59.6	79.3$^{+19.7}$	84.4$^{+24.8}$	64.7	83.7$^{+19.0}$	84.5	77.9	89.6$^{+11.7}$
SVHN	13.4	27.7$^{+14.3}$	38.8$^{+25.4}$	31.3	38.5$^{+7.2}$	36.2	57.0	38.0$^{-19.0}$
FER2013	41.4	43.0$^{+1.6}$	48.1$^{+6.7}$	46.3	43.2$^{-3.1}$	44.5	50.1	50.3$^{+0.2}$
RenderedSST2	58.6	52.3$^{-6.3}$	54.3$^{-4.3}$	60.5	54.4$^{-6.1}$	57.9	68.9	56.0$^{-12.9}$
Pets	87.3	86.9$^{-0.4}$	89.2$^{+1.9}$	89.0	89.2$^{+0.2}$	90.3	93.3	91.9$^{-1.4}$
Caltech-101	81.6	83.2$^{+1.6}$	83.1$^{+1.5}$	82.2	83.6$^{+1.4}$	83.2	83.3	84.0$^{+0.7}$
VOC2007-Cl	76.4	75.8$^{-0.6}$	78.8$^{+2.4}$	78.3	76.8$^{-1.5}$	76.4	78.3	75.6$^{-2.7}$
SUN397	62.5	67.0$^{+4.5}$	68.5$^{+6.0}$	64.4	69.6$^{+5.2}$	69.8	67.6	72.6$^{+5.0}$
FGVC Aircraft	19.6	16.7$^{-2.9}$	23.1$^{+3.5}$	24.3	17.7$^{-6.6}$	18.5	31.8	25.0$^{-6.8}$
Country211	17.2	14.8$^{-2.4}$	16.5$^{-0.7}$	22.8	18.1$^{-4.7}$	18.9	31.9	23.0$^{-8.9}$
DTD	44.3	54.6$^{+10.3}$	53.9$^{+9.6}$	44.9	51.3$^{+6.4}$	55.5	55.3	60.5$^{+5.2}$
GTSRB	32.6	42.0$^{+9.4}$	36.5$^{+3.9}$	43.3	43.5$^{+0.2}$	49.4	50.6	49.9$^{-0.7}$
STL10	97.1	95.6$^{-1.5}$	96.5$^{-0.6}$	98.2	97.0$^{-1.2}$	97.0	99.4	98.1$^{-1.3}$
Retino	45.5	24.2$^{-21.3}$	19.1$^{-26.4}$	3.3	7.4$^{+4.1}$	9.2	73.3	6.0$^{-67.3}$
EuroSAT	50.4	51.5$^{+1.1}$	50.3$^{-0.1}$	55.9	50.3$^{-5.6}$	58.2	62.6	62.3$^{-0.3}$
RESISC45	53.6	54.5$^{+0.9}$	61.9$^{+8.3}$	58.2	58.5$^{+0.3}$	61.4	63.4	67.4$^{+4.0}$
PCAM	62.3	55.9$^{-6.4}$	50.7$^{-11.6}$	50.7	59.6$^{+8.9}$	55.2	52.0	49.6$^{-2.4}$
CLEVR Counts	23.2	16.2$^{-7.0}$	19.2$^{-4.0}$	21.2	28.7$^{+7.5}$	23.9	19.4	24.2$^{+4.8}$
CLEVR Dist	16.3	15.9$^{-0.4}$	16.8$^{+0.5}$	15.8	24.5$^{+8.7}$	15.9	16.1	14.9$^{-1.2}$
DSPRITES Orient	2.4	1.9$^{-0.5}$	2.3$^{-0.1}$	2.3	2.9$^{+0.6}$	2.7	2.3	2.6$^{+0.3}$
DSPRITES pos	3.6	2.8$^{-0.8}$	3.1$^{-0.5}$	3.0	3.2$^{+0.2}$	4.3	3.2	3.0$^{-0.2}$
SmallNORB Elv	12.7	9.9$^{-2.8}$	11.0$^{-1.7}$	12.2	10.0$^{-2.2}$	11.0	11.5	11.0$^{-0.5}$
SmallNORB Azim	6.1	4.5$^{-1.6}$	5.2$^{-0.9}$	5.2	6.0$^{+0.8}$	5.5	4.5	5.3$^{+0.8}$
DMLAB	19.3	17.3$^{-2.0}$	18.9$^{-0.4}$	15.5	15.1$^{-0.4}$	14.8	16.3	18.7$^{+2.4}$
KITTI Dist	27.4	28.8$^{+1.4}$	17.6$^{-9.8}$	26.4	18.1$^{-8.3}$	28.1	21.8	20.1$^{-1.7}$
VTAB+(Avg.)	45.4	45.6$^{+0.2}$	47.9$^{+2.5}$	47.5	48.3$^{+0.8}$	49.2	55.7	51.8$^{-3.9}$

Table 7: Comparison between CLIP models trained on LAION (400M, 2B) and the original CLIP models [58] trained on OpenAI's WebImageText (WIT) dataset. We show zero-shot top-1 classification accuracy (%) on the 35 datasets that are part of VTAB+. We highlight the difference (+/-) between LAION models and original CLIP WIT models for each model size (except B/16+, for which there is no CLIP WIT checkpoint).

Model	Pre-training	Flickr30K (1K test set)					
		Image → Text			Text → Image		
		R@1	R@5	R@10	R@1	R@5	R@10
ViT-B/32	CLIP WIT	77.5	94.7	98.2	58.8	83.3	89.7
	LAION-400M	78.9	94.0	97.1	61.7	85.5	90.9
	LAION-2B-en	84.3	96.3	98.4	66.3	88.2	93.2
ViT-B/16	CLIP WIT	81.9	96.2	98.8	81.9	96.2	98.8
	LAION-400M	83.3	96.8	98.5	65.5	88.3	93.0
ViT-B/16+	LAION-400M	86.5	97.1	98.8	68.0	88.9	94.0
ViT-L/14	CLIP WIT	85.1	97.3	99.0	65.2	87.3	92.0
	LAION-400M	87.6	97.7	99.5	70.3	90.9	94.6

Table 8: CLIP Zero-Shot retrieval results on the Flickr30K test set. We show retrieval performance at 1, 5, and 10 samples for both image to text and text to image.

Model	Pre-training	MSCOCO (5K test set)					
		Image → Text			Text → Image		
		R@1	R@5	R@10	R@1	R@5	R@10
ViT-B/32	CLIP WIT	50.0	75.0	83.3	30.4	54.8	66.1
	LAION-400M	53.5	77.2	85.4	34.9	60.3	71.1
	LAION-2B-en	56.4	79.6	87.4	38.7	64.1	74.4
ViT-B/16	CLIP WIT	51.7	76.8	84.3	32.7	57.8	68.2
	LAION-400M	56.5	80.4	87.3	37.9	63.2	73.3
ViT-B/16+	LAION-400M	58.6	81.6	88.4	40.0	65.5	75.1
ViT-L/14	CLIP WIT	56.0	79.5	86.9	35.3	60.0	70.2
	LAION-400M	59.3	81.9	89.0	42.0	67.2	76.6

Table 9: CLIP Zero-Shot retrieval results on the MSCOCO test set. We show retrieval performance at 1, 5, and 10 samples for both image to text and text to image.

Prompt: A couple of bananas hanging from a metal hook.

GLIDE

LAIONIDE-v3

Prompt: A group of people that are standing in the street.

GLIDE

LAIONIDE-v3

Prompt: A street scene with focus on a bicycle under a window.

GLIDE

LAIONIDE-v3

Figure 13: **Comparison of GLIDE and LAIONIDE-v3 Generations.** We compare the output of GLIDE and our LAIONIDE-v3 across three different prompts. The top row of each section depicts GLIDE's results, while the bottom row depicts LAIONIDE-v3's resutls.

F.2 Stable Diffusion

Stable Diffusion is a generative latent diffusion model trained on various LAION-5B subsets:

- 237,000 steps at 256x256 on LAION-2B-en
- 194,000 steps at 512x512 on laion-high-resolution
- 515,000 steps at 512x512 on laion-improved-aesthetics
- 390,000 steps at 512x512 on laion-improved-aesthetics with 10% dropping of the text conditioning

Here we show representative generated samples for an artistic (Fig. 14) and a photorealistic (Fig. 15) image. For more technical details, we refer to the Stable Diffusion github repository[34].

Figure 14: **"The sigil of water by Gerardo Dottori, oil on canvas"**
Generated by Stable Diffusion

G Further Discussion on Safety and Ethics

G.1 Privacy

As any other dataset of links obtained from Common Crawl that gathers content from publicly available Internet, LAION-5B can contain links to images with personal information, like to photos of faces, medical images or other personal related content. Tools like CLIP retrieval (see Appendix Section C.4 for more details) provided by LAION make it possible for the users to find out by text or image prompt whether any of the links crawled for LAION-5B point to their personal data and if yes, where on the public internet the corresponding data is hosted. Thus, for the first time, the

[34] `https://github.com/CompVis/stable-diffusion/`

Figure 15: **"A wide river in the jungle, Provia, Velvia"**
Generated by Stable Diffusion

broad public can take a look inside of a typical large-scale crawled dataset and become aware of the possible content of datasets that can be used for model training. As most of institutions and companies use same crawling procedures to obtain their closed datasets, we thus also hope to increase awareness for the risks which publicly available data can be used and exploited by third parties who do not disclose their data collection and application procedures. At the same time, researchers can access LAION-5B to study privacy related issues in such data and develop measures that increase safety of applications arising from training models on data crawled from public internet.

As LAION tools empower people to discover problematic personal or copyrighted content available in the public internet, the users can also initiate procedures of removing corresponding images from the public internet by contacting the responsible host providers that have published those images following the links provided in LAION-5B. In addition, we also provide a contact form on our website [35] where requests for removal or blacklisting of the corresponding links from LAION-5B can be processed.

Further, to mitigate privacy concerns, there exist methods that allow personal human attributes like faces to be obfuscated [87] or generated [48] and thus made anonymous, without hurting the quality and richness of learned representations. Especially generation based methods can be applied to open data like LAION-5B to create training datasets that do not contain any private facial data, while still allowing to learn proper face representation during training. This line of work is currently in progress in LAION community.

[35] https://laion.ai/dataset-requests/

G.2 Potential Biases Induced by CLIP Filtering

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 [6], 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.

Author contributions

- **Christoph Schuhmann**: He led this project and built POCs for most of its components including clip filtering, the safety model, the watermark model and the BLIP inference tuning project.
- **Richard Vencu**: System architecture and download script optimizations, GPU assisted filtering. Set up the AWS infrastructure.
- **Romain Beaumont**: Guidance on scaling for the Common Crawl filtering pipeline. Built and ran the dataset preparation pipeline: pyspark deduplication job, img2dataset, CLIP inference, autofaiss, safety tags.
- **Clayton Mullis**: DALLE-pytorch training/analysis, WDS filtering, trained generative models (LAIONIDE) using LAION-5B.
- **Ludwig Schmidt**: Provided advice on experiment design, scaling, ethical and social content, and paper writing.
- **Jenia Jitsev**: scientific organization & manuscript writing, ethical and social content, experiments planning and design, compute and storage resource acquisition, general supervision.
- **Robert Kaczmarczyk**: Established WDS architecture, performed DALL-E training runs, balancing calculation, sample (NSFW, watermark, caption quality) annotation, manuscript writing coordination, supervision and revision.
- **Theo Coombes**: He was one of our first contributors & build the first versions of our worker swarm system. Without his enthusiasm this project might never have taken off.
- **Aarush Katta**: Trained the watermark model.
- **Cade Gordon**: Ran distributed inference for the watermark tags, trained the CLIP models on JUWELS Booster, and led the paper writing.
- **Mehdi Cherti**: Evaluated the CLIP-B/32, B/16, B/16+ and L/14 model, performed debugging of distributed training, executed experiments on JUWELS Booster, performed results collection, distillation and analysis, manuscript writing.
- **Ross Wightman**: Ross debugged & trained the CLIP-B/32, B/16, B/16+ and L/14 model and executed experiments on JUWELS Booster.
- **Katherine Crowson**: Contributed to development of latent diffusion and stable diffusion. Fine-tuned generative models on subsets of LAION-5B.
- **Patrick Schramowski**: Patrick helped with NSFW and otherwise inappropriate content tagging. Further, he wrote the corresponding parts as well as the ethical and social content.
- **Srivatsa Kundurthy**: Co-wrote the datasheet, researched usage cases & related works, trained face classifier and developed visualizations.
- **Mitchell Wortsman** Initially created openCLIP, provided insights on scaling, performed experiments evaluating few-shot fine-tuning performance and robustness on ImageNet and other downstream datasets

Acknowledgments details

We want to thank our open community for their continuous efforts for openly available datasets and models. Without the broad support from the community, especially in the early crawling days with decentralized compute support, this project would not have been possible.

Moreover, the following organizations and persons contributed to this project:

- **Aran Komatsuzaki**: He led the initial crawling@home image-text-pair dataset building project (the predecessor of LAION-400M).
- **Andreas Köpf**: He conducted the hyperparameter search for the inference strategies with the BLIP image-captioning model.
- **Bokai Yu**: Accomplished most of the work to make the knn index building tool autofaiss work in a distributed setting.
- **John David Pressman**: Provided aestethic dataset for creating aestethic LAION subset to fine-tune GLIDE.
- **Natalie Parde** Assisted in manuscript revisions.
- **Gabriel Ilharco** Initially created OpenCLIP and gave valuable insights on scaling.
- **Fredde Frallan** Provided zero-shot retrieval results.
- **Hugging Face**: provided financial and computing support, helped hosting LAION-5B as well as related subsets.
- **Emad Mostaque (Stability AI)**: provided financial and computation support for open-source datasets and models.