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 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).*

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

- 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.

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.

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

[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

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 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 *movie*s, *game*s and *comic*s are potentially tagged, depending on the displayed content. Further filtering depends highly on the use-case and users' opinions.

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.

[28] https://www.wikiart.org

[29] https://www.gwern.net/Danbooru2021

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

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

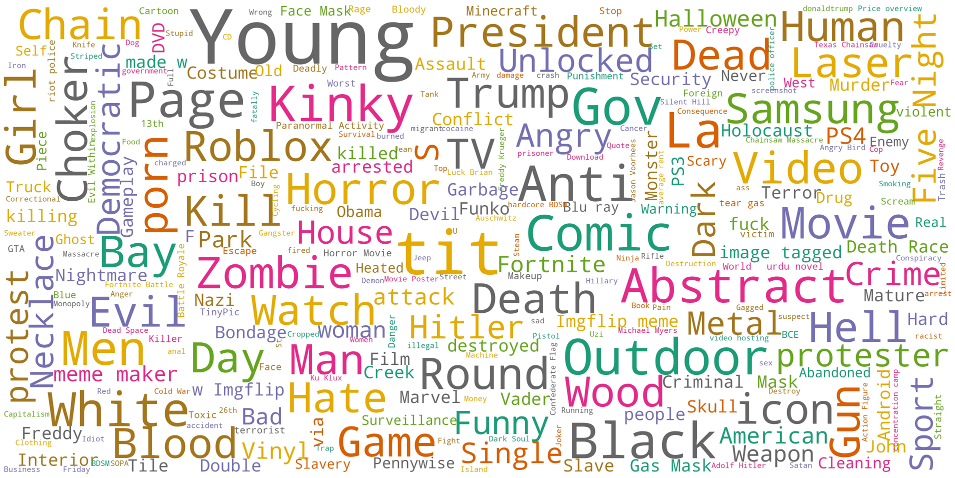


Figure 5: Word cloud based on [68] documenting the potentially inappropriate image content of the LAION-5B subset which contains text in English language. Provided alternative text is used as text description of the images. Word size is proportional to the word counts and rank in descriptions corresponding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

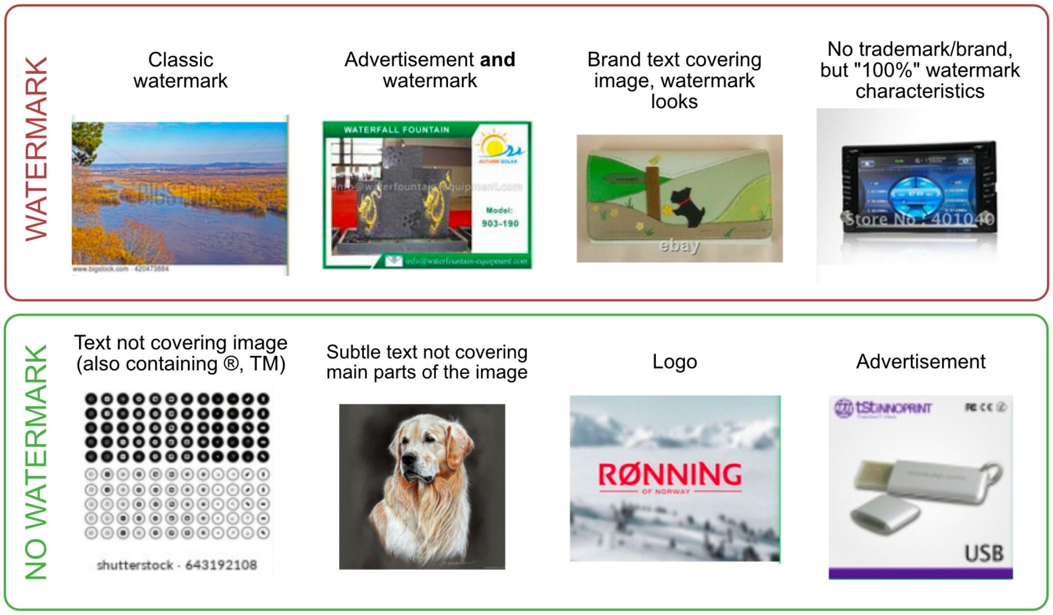
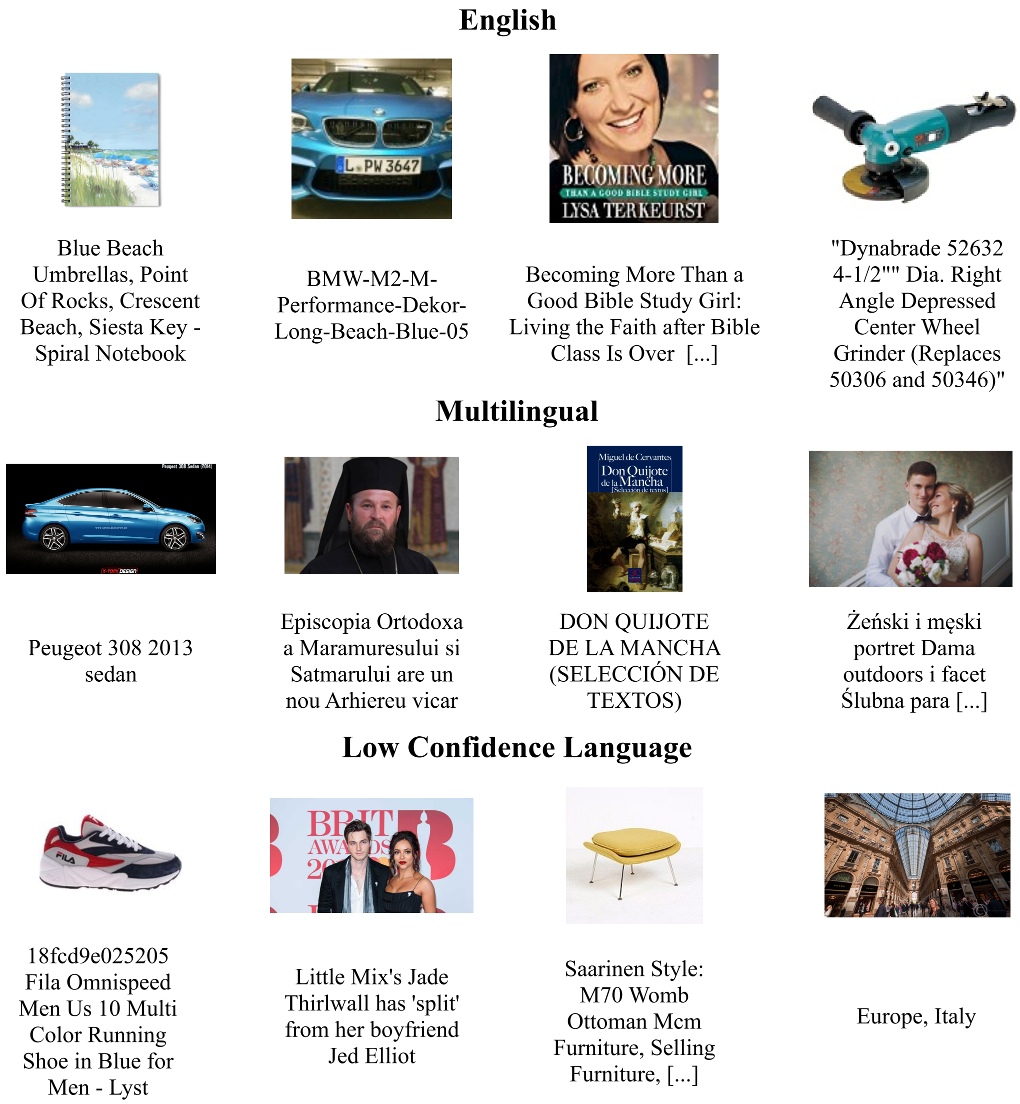

WATERMARK
Classic watermark
Advertisement and watermark
Brand text covering image, watermark looks
No trademark/brand, but "100%" watermark characteristics
bigstock
www.bigstock.com · 420473884
WATERFALL FOUNTAIN
AUTUMN SOLAR
Model: 903-190
info@waterfountain-equipment.com
ebay
Store No: 401040
NO WATERMARK
Text not covering image (also containing ®, TM)
Subtle text not covering main parts of the image
Logo
Advertisement
shutterstock · 643192108
RØNNING
OF NORWAY
tStINNOPRINT
FC CE
USB

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

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

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.



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.

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.

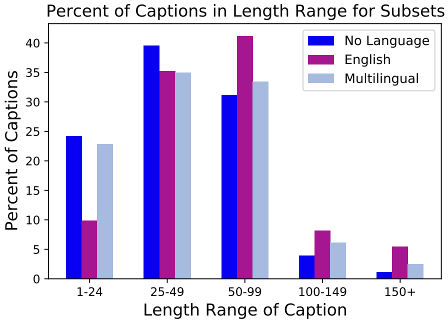


Figure 8: **Caption Character Length.** Each of the LAION-5B subsets contains similar frequencies and exhibit a right skew.

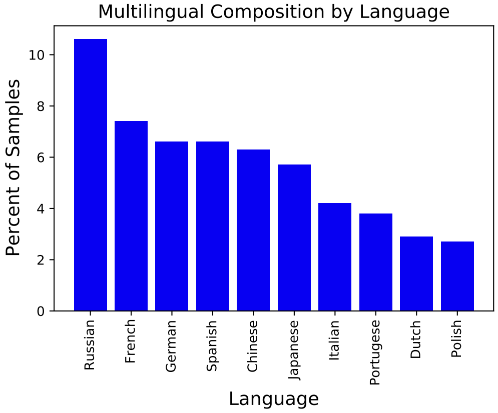


Figure 9: **Multilingual Language Frequency.** The 10 most frequent languages seem to be largely of European and East Asian origin.

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].

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+).

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 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.

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

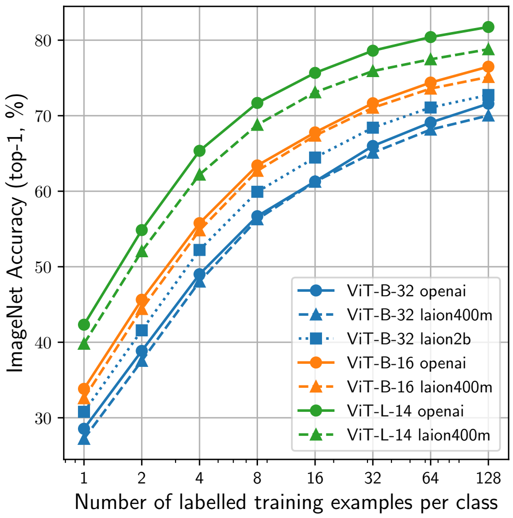


Figure 10: Evaluating few-shot linear probe performance on ImageNet. We evaluate i) models trained on various LAION subsets and ii) the original CLIP models. Models trained on LAION show similar transfer performance to those trained by OpenAI. Also evident is clear effect of model or data scale on transfer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

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.

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).

	B/32			B/16		B/16+	L/14	
Dataset	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).

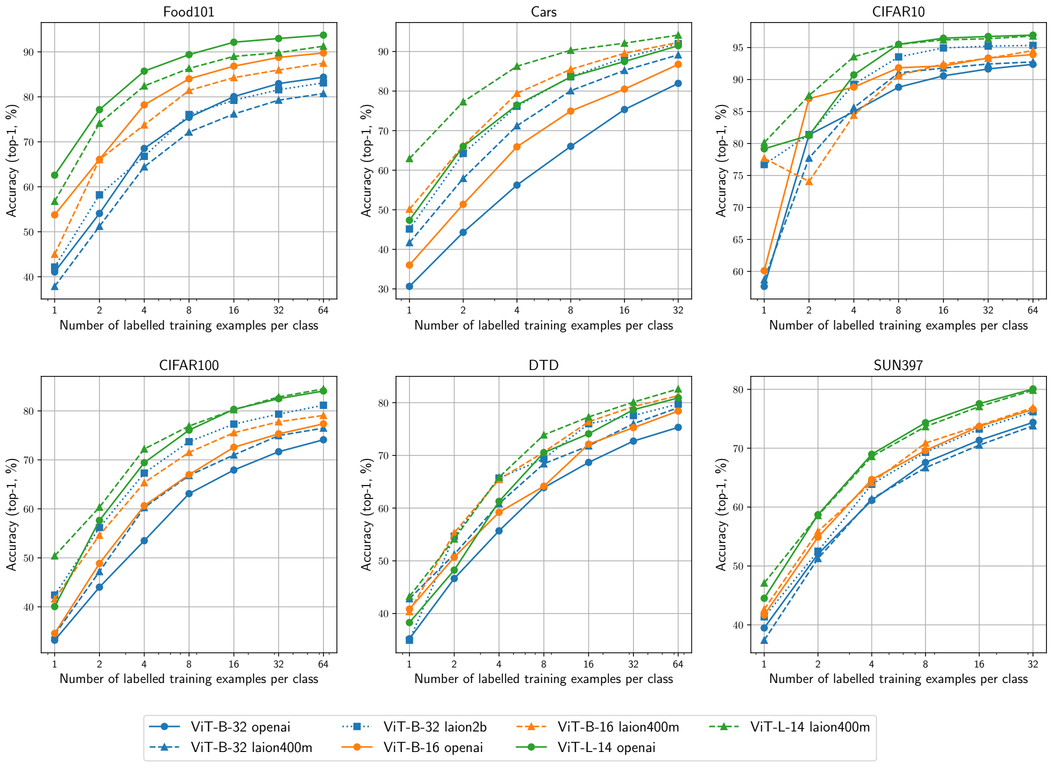


Figure 11: Evaluating few-shot linear probe performance on 6 datasets commonly used to benchmark transfer [34]. We evaluate i) models trained on various LAION subsets and ii) the original CLIP models. We evaluate performance on Food101 [7], Cars [35], CIFAR-10 & 100 [37], DTD [12] and SUN397 [85].

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 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.

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.

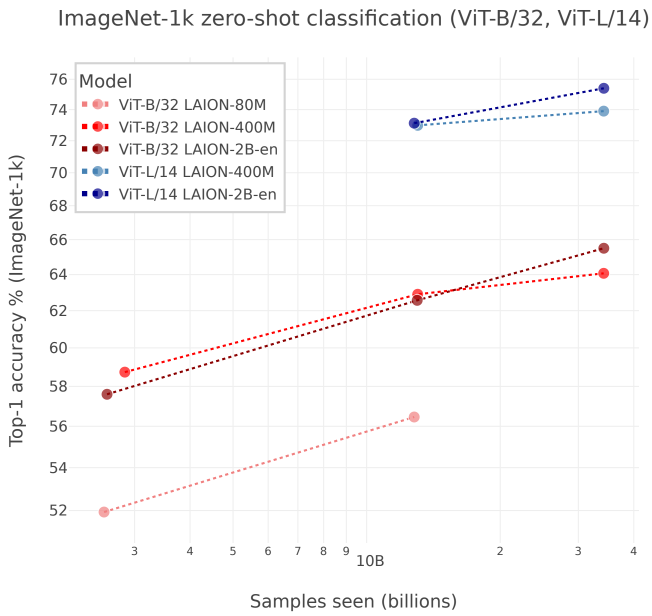


Figure 12: ViT-B/32 and ViT-L/14 additional experiments where we vary the amount compute (3B, 13B, and 34B images seen) and LAION subset size (80M, 400M, 2B). We evaluate the models on zero-shot Imagenet-1k classification. Seeing same number of samples on larger data scale leads consistently to better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

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

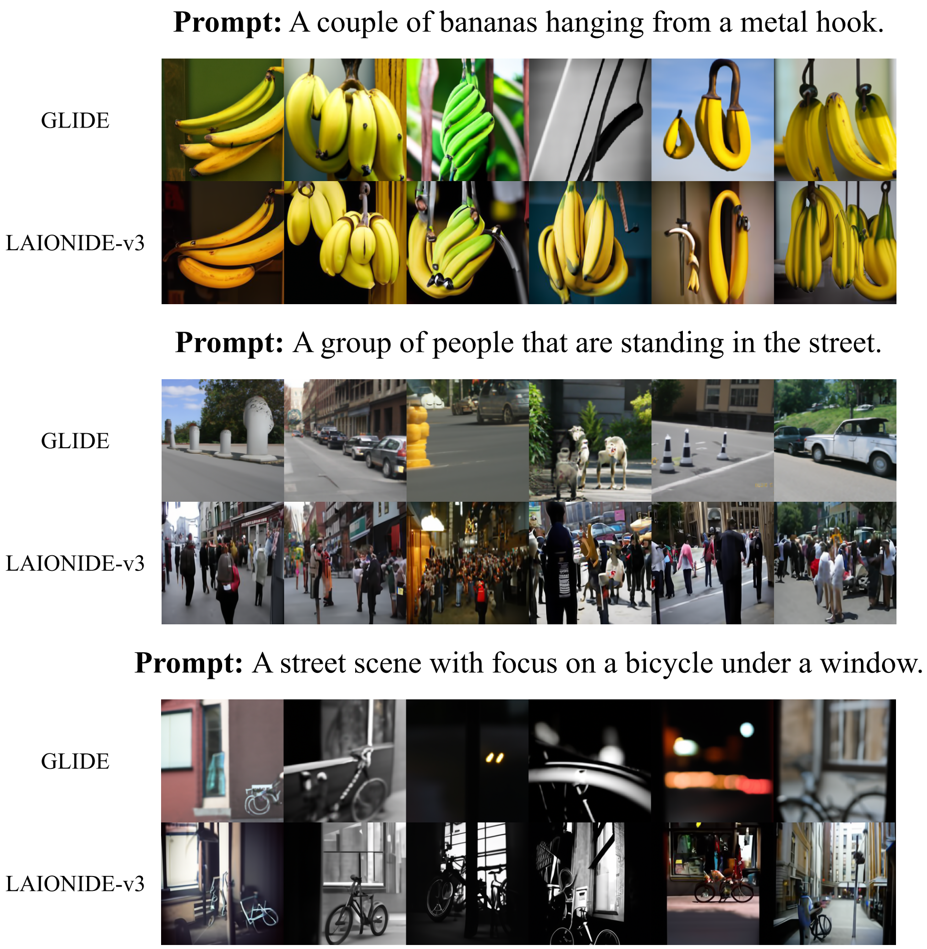
OpenAI released checkpoints for the GLIDE [52] architecture to the public, but only released checkpoints trained on a filtered dataset removing hate-symbols and humans. These models can do a lot, but are incapable of generating imagery of humans. To evaluate the LAION dataset and its generalization capabilities,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

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



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

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

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

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.

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`

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].

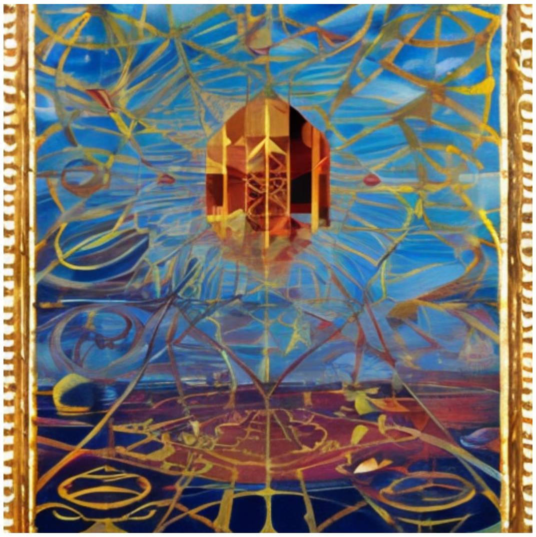


Figure 14: **"The sigil of water by Gerardo Dottori, oil on canvas"**
Generated by Stable Diffusion

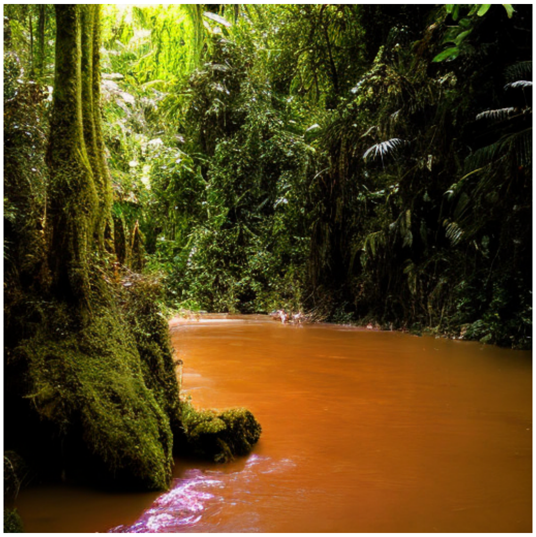


Figure 15: **"A wide river in the jungle, Provia, Velvia"**
Generated by Stable Diffusion

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

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

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 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.

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.

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

		Flickr30K (1K test set)					
Model	Pre-training	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.

		MSCOCO (5K test set)					
Model	Pre-training	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.

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 aesthetic dataset for creating aesthetic 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.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes] We discuss technical limitations in section 6 and discuss the safety and ethics of the work in section 7.
   (c) Did you discuss any potential negative societal impacts of your work? [Yes] We discuss this in section 7
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [N/A]
   (b) Did you include complete proofs of all theoretical results? [N/A]
3. If you ran experiments (e.g. for benchmarks)...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] We include links throughout the whole work.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] See tables 3 and 4.
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [No] Due to the computational cost of the models, we were unable to produce error bars.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes] See table 4.
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [Yes]
   (c) Did you include any new assets either in the supplemental material or as a URL? [Yes]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [Yes] See the datasheet in appendix section A.
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [Yes] See the datasheet in appendix section A.
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]