FairDeDup: Detecting and Mitigating Vision-Language Fairness Disparities in Semantic Dataset Deduplication

Eric Slyman[1,2*] Stefan Lee[1] Scott Cohen[2] Kushal Kafle[2]

[1]Department of EECS, Oregon State University [2]Adobe Research

Abstract

Recent dataset deduplication techniques have demonstrated that content-aware dataset pruning can dramatically reduce the cost of training Vision-Language Pretrained (VLP) models without significant performance losses compared to training on the original dataset. These results have been based on pruning commonly used image-caption datasets collected from the web – datasets that are known to harbor harmful social biases that may then be codified in trained models. In this work, we evaluate how deduplication affects the prevalence of these biases in the resulting trained models and introduce an easy-to-implement modification to the recent SemDeDup algorithm that can reduce the negative effects that we observe. When examining CLIP-style models trained on deduplicated variants of LAION-400M, we find our proposed FairDeDup algorithm consistently leads to improved fairness metrics over SemDeDup on the FairFace and FACET datasets while maintaining zero-shot performance on CLIP benchmarks.

1. Introduction

Recent Vision-Language Pretrained (VLP) models [55] that learn to align image and language encodings have demonstrated strong zero-shot performance on many standard perception tasks [12, 16, 71, 73]. Beyond these, VLP models have enabled complex downstream applications ranging from visually-aware chatbots [42, 44] and language-based image segmentation [37, 79] to instruction-guided robotics [62, 69] and semantic mapping of 3D scenes [35, 61]. The rapid adoption and widespread impact of these models is due in part to the incredibly broad range of content they can represent effectively – a scope far exceeding prior models trained on manually-curated, closed-world datasets [16, 43]. To acquire this capability, VLP models are trained on massive open-world datasets of image-caption pairs collected from the internet [58]. VLP models improve reliably with additional training data [13], driving the number of examples in these datasets into the billions [59]. This scale of uncurated data introduces at least two challenges – 1) training can be extremely costly, and 2) manual data curation to reduce undesirable social biases is economically prohibitive. In this work, we explore how dataset deduplication techniques developed to reduce training costs may exacerbate or ameliorate these biases in trained models.

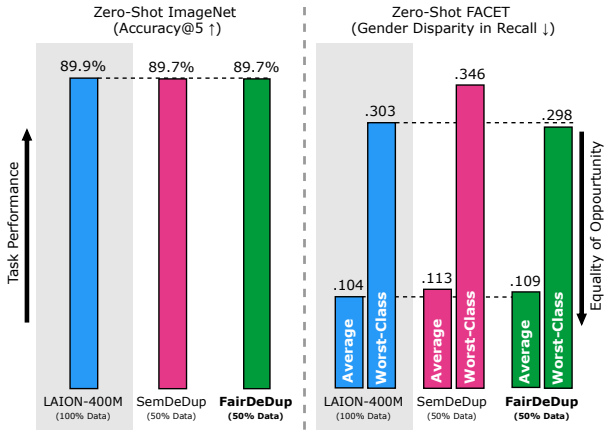

Figure 1. Training models on deduplicated data can yield similar results to the full-data setting on standard tasks like zero-shot ImageNet [16] classification (**left**, higher is better ↑). However, impacts on subgroup performance have not been studied. We discover cases such as gender disparity (**right**, lower is better ↓) where deduplication reinforces existing biases on FACET [26]. FairDeDup preserves performance while reducing bias from deduplication and, in some cases, w.r.t. the full-data setting.

While larger pretraining datasets generally yield better model performance [13], the massive web-scraped datasets commonly used for training VLP models contain many identical samples (duplicates) or samples that capture nearly the same content under similar imaging conditions (semantic duplicates [1]). Several recently developed techniques for data pruning/deduplication have demonstrated that aggressive removal of these duplicates has limited im-

*Work conducted during Slyman's 2023 summer internship at Adobe.

pact on the task performance of trained models [1, 48, 63]. For example, Abbas et al. [1] found that pruning LAION-400M [58] by 50% resulted in trained models that achieved average performance within 0.5% of their full-data analogs across a range of common benchmark tasks – effectively cutting training time in half.

However, these web-scale datasets contain a plethora of problematic social biases and harmful stereotypes [5, 6, 23]. These biases can often then be reflected in the behavior of models trained on these datasets [2, 27, 28, 68]. To better understand and reduce these potential harms, there is increased interest in analyzing the composition of these datasets and their downstream effects on trained models [5, 23, 77]. Deduplication techniques introduce another algorithmic step between dataset and model training that may systematically alter the data distribution – potentially amplifying, maintaining, or reducing the effect of dataset biases. Given that deduplication techniques will likely be widely deployed as cost-saving measures, understanding how their design affects the behavior of downstream models in terms of bias and fairness is a timely but unexamined question.

To study this question, we investigate the fairness outcomes of CLIP-style [55] VLP models trained on the LAION-400M dataset [58] pruned with SemDeDup [1]. Replicating the results of Abbas et al. [1], we find task performance on CLIP Benchmark [12] is only marginally affected; however, evaluation on the fairness-focused FairFace [34] and FACET [26] datasets suggest deduplication results in mixed effects compared to the full-data setting. We observe increased disparities across gender, but both positive and negative changes for disparities across skin tone and age. Based on these findings, we propose FairDeDup – a fairness-aware data pruning algorithm that makes pruning decisions to improve representation of specified *sensitive concepts* (*e.g.*, *gender*, shown in Fig. 1). The implementation of FairDeDup is a simple modification to SemDeDup and specifying concepts can be done in natural language. Our large-scale experiments show that FairDeDup leads to improved fairness outcomes comapred to SemDeDup while maintaining comparable performance on standard zero-shot and retrieval-based performance benchmarks. To better understand the deduplication process, we run a smaller scale study deduplicating demographic-labeled data – finding that FairDeDup consistently retains more images depicting minority classes than SemDeDup.

Contributions. We summarize our contributions below:

- We conduct, to our knowledge, the first large-scale experiment evaluating the fairness outcomes of training large-scale vision language models on pruned data – training CLIP-style models on full and deduplicated versions of the popular LAION-400M dataset then evaluating on standard fairness benchmarks for VLP models.
- We find that models trained on SemDeDup [1] pruned data have varied effects on fairness outcomes from the full-data model; reinforcing some biases and mitigating others.
- We introduce FairDeDup, a simple and efficient modification to SemDeDup that improves fairness outcomes while retaining task performance – improving fairness outcomes over SemDeDup in nearly all cases studied.

2. Related Work

Vision-Language Fairness. Vision and language models have been shown to learn, reflect, and amplify problematic social biases. For example, vision systems have been shown to dehumanize minority groups by identifying them as animals [19] and degrade in task performance on intersectional combinations of gender and skin tone [9]. Likewise, language models are known to learn gendered associations of professions [8], increase sentiment-intensity along racial lines [38], and a myriad of other problems documented in [7, 65]. Vision-language models are not exempt from these problems [27, 47, 52] and can even reinforce them [64, 77].

Contemporary Vision-Language Pretrained models are frequently pretrained on massive but uncurated data scraped from the internet [11, 32, 40, 55]. While web-scale data is shown to improve performance, it also teaches models "*misogyny, pornography, and malignant stereotypes*" [5]. VLP models demonstrate dehumanizing behavior with respect to racial subgroups in zero-shot text-image retrieval [2, 3], show bias related to gender [23, 27, 28, 68], age [23] and skin tone [23, 28, 68, 76] in image captioning, and also demonstrate biases relating to age, gender, skin tone, and ethnicity in text-image retrieval [23, 78]. These behaviors are attributed to the use of uncurated web-scale datasets in pretraining VLP models [5, 6, 23].

Mitigations for bias in VLP models typically include fairness-aware training [75] or post-hoc methods to disentangle useful concepts from sensitive attributes [3, 14, 60]. Unlike these methods, we seek to prevent bias from being reinforced in the dataset, rather than removing bias from the model itself. Though early vision-language fairness literature frequently calculates WEAT [10] and SEAT [49] embedding association measures extended for the multimodal setting [31, 57], these measures have been shown to be overly sensitive to small changes in model architecture and outputs [3]. As such, VLP model fairness is primarily evaluated on CelebA [45] and FairFace [34]. Recent datasets such as PHASE [23] and FACET [26] allow for the study of bias on "in the wild" data across diverse subgroups.

Dataset Pruning. Several techniques exist for reducing the size of a dataset while preserving, or even improving, performance. We consider all techniques under this umbrella as *dataset pruning* algorithms. Coreset selection chooses a weighted subset of training samples which closely estimate the full dataset's gradient [25, 50] to perform data-

efficient training with little loss in performance. However, these methods do not scale well with dataset size and frequently require class labels [63]. The most similar work to ours among coreset selection algorithms is the recent D2 Pruning [48]. D2 Pruning utilizes graph based methods to select samples that are both *hard* and *diverse* across a data distribution. While promising, D2 Pruning does not evaluate any fairness outcomes and is only demonstrated to scale to DataComp Small (12.8M) [22], a low accuracy setting for VLP models peaking around 5% top-1 zero-shot ImageNet [16] accuracy. In comparison, base-sized CLIP-style models can range from 67-74% accuracy with web-scale data on the same task. We refer readers to [54] for a more thorough review of coreset selection algorithms.

Large-scale deduplication typically attempts to find exact perceptual duplicates using techniques like perceptual hashing [20] or filtering [22] on image-text CLIP scores and target classes (*e.g.*, filtering to images close to ImageNet classes). Abbas et al. [1] introduces the concept of semantic duplicates, images with similar semantic meaning that are not perceptually the same image, alongside SemDeDup, a formalized version of the unsupervised deduplication algorithm from [63]. SemDeDup has been shown to be capable of significantly reducing dataset size with only marginal impact on performance. We choose to study SemDeDup due to the ubiquity of its underlying selection method among contemporary deduplication algorithms - cosine similarity between samples - and scalable nature. To our knowledge, we are the first to study the effect of data pruning on the fairness outcomes of VLP models and study the effects of fairness-aware pruning on their behavior.

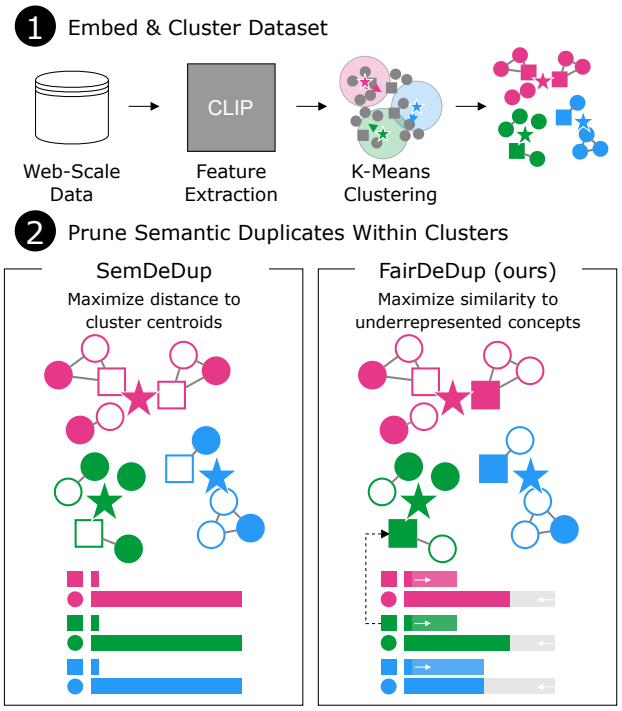

Figure 2. The semantic deduplication pipeline following three clusters (★,★,★) with two subgroups (■,●). Connected shapes are duplicates. We **(1) embed** all images from the dataset with a pretrained model then partition with k-means to enable efficient search during **(2) deduplication**. We make a simple modification to the maximum distance selection heuristic used by Abbas et al. [1] **(left)** to improve subgroup diversity by preserving samples which maximize similarity to poorly represented sensitive concepts according to user-specified concept prototypes (**right**).

3. FairDeDup: Fair Semantic Deduplication

There frequently exists sensitive attributes in data for which it is desirable to obtain some notation of fairness [21]. For example, we may seek *demographic parity* for *gender* so that individuals do not receive differing treatment based on their gender identity. Such outcomes are usually based on social norms, organizational ethics, or even codified into discrimination law [4, 15, 29, 53]. Our goal is to improve post-deduplication fairness outcomes concerning these sensitive groups. To achieve this, we propose boosting the representation of underrepresented sensitive subgroups on the internet (*e.g.*, women of color) in the post-pruning dataset distribution. We allow for user-defined natural language *sensitive concepts*, which captures these subgroups for consideration in the deduplication process, and leverage them to bias the selection of preserved samples towards those concepts which are currently underrepresented.

3.1. Preliminaries: SemDeDup

We implement FairDeDup as a lightweight modification to the SemDeDup algorithm, which we describe here for completeness. Abbas et al. [1] identify that pruning both exact perceptual duplicates (*e.g.*, copies of the same image) and those that carry redundant semantic information (*e.g.*, many photos of the same object from differing angles), denoted semantic duplicates, is helpful for improving the data efficiency of training large models. To achieve this, they propose SemDeDup [1], an extension of the unsupervised pruning metric from Sorscher et al. [63] to web-scale data.

To identify duplicates, SemDeDup first leverages pretrained foundation models (*e.g.* CLIP [55]) to embed all images in the dataset into a semantically meaningful feature space. Naïvely thresholding embedding similarity between all points to detect duplicates requires $\mathcal{O}(n^2)$ pairwise comparisons and is intractable for web-scale data like LAION-400M, which requires computing $\approx 1.5 \times 10^{17}$ cosine similarities. To mitigate this, the dataset is partitioned using an efficient K-means algorithm under the assumption that pairwise similarity need only be calculated for approximately similar samples. SemDeDup then considers the resulting $\mathcal{O}(n^2/k)$ pairwise similarities on an independent per

cluster basis. Within each cluster, they determine sets of samples within a $1-\epsilon$ similarity threshold as duplicates and keep only the sample most distant from the cluster centroid. While this selection heuristic is motivated by the hardness hypothesis of Sorscher et al. [63], ablations show that the algorithm is robust to choosing even a random sample.

3.2. FairDeDup

Due to the robustness of SemDeDup to the choice of selection heuristic on performance, we seek instead to replace the heuristic with one that can support our fairness motivation. We provide an overview following shared and unique steps of SemDeDup and FairDeDup in Fig. 2.

Sensitive Concept Prototypes. Given a list of user-defined sensitive concepts C that are desired to be represented in the pruned dataset, we denote the *concept prototype* P_i for a sensitive concept $C_i \in C$ as the average text embedding of the set of captions generated from template strings (*e.g.*, "*A photo of a* $\{C_i\}$") capturing that concept. As is common for VLP models, we assume the embedding model supporting image clustering can also produce image-text alignment scores [32, 41, 55, 72, 74] and consider the case where alignment is determined as the cosine similarity between the representations produced by a vision $\Phi_I : I \to \mathbb{R}^d$ and text $\Phi_T : T \to \mathbb{R}^d$ encoder:

$$sim(I, T) = \Phi_I(I)^T \Phi_T(T) \; / \; \|\Phi_I(I)\| \|\Phi_T(T)\|. \quad (1)$$

We measure how well an image aligns with a sensitive concept by measuring the image-text similarity between that image and the concept prototype $sim(I, P_i)$. We choose concepts that both relate to commonly protected demographic subgroups of people and are annotated in common fairness datasets, such as ones based in race and gender. Additional details on the selection of sensitive concepts and a list of all concepts used are given in the appendix. While this work focuses on text-based prototypes, we note that our methodology trivially extends to image-based ones and beyond, as described in Sec. 6.

Sample Preservation Heuristic. To determine which samples to prune, we consider duplicate *neighborhoods*: the set of images within $1-\epsilon$ similarity of a given point, and preserve only one sample from each neighborhood. For each cluster produced by k-means, we track the running average similarity between preserved samples in that cluster and the sensitive concept prototypes. Until all samples are visited, we randomly select an unvisited sample, calculate the similarity between all samples in its neighborhood and the prototypes, and keep only the sample that maximizes similarity to the least similar running average prototype, marking all points in the neighborhood as visited. We preserve the sample with the highest average similarity across concept prototypes for the first neighborhood visited in a cluster.

```
# Input: prototypes, embeddings, eps
# Get similarity with concept prototypes
proto = embeddings @ prototypes.T

balance = AverageMeter(prototype.shape[0])
tovisit = torch.ones(embeddings.shape[0])
while tovist.any():
    # Find an unvisited neighborhood
    node = torch.where(tovisit)[0][0]
    sims = embeddings[node] @ embeddings.T
    neighbors = torch.where(sims > 1 - eps)[0]

    # Maximize least represented concept
    c = balance.get_min_concept()
    point = proto[neigbors][:, c].argmax()
    balance.update(point)

    log_and_keep(point)
    tovisit[neighbors] = 0
```

Figure 3. PyTorch-style pseudo-code for FairDeDup selection given concept `prototypes`, within cluster `embeddings`, and an `eps` similarity threshold for determining neighborhoods. We omit the base case where the first sample selected within a cluster is the one with the highest average concept prototype similarity.

We track running average similarity on a per cluster basis for two reasons: 1) to avoid a synchronous update step between workers processing clusters in parallel and 2) to prevent algorithmic "gaming" of the selection criteria by balancing concept representation on clusters which highly represent a concept due to some stereotyped notion. Given two clusters primarily composed of doctors and nurses, for example, per cluster processing prevents balancing underselection of feminine presenting doctors by overselecting feminine presenting nurses. We provide pseudo-code for the FairDeDup selection heuristic in Fig. 3. We visualize random samples after pruning a cluster manually identified to be primarily composed of people with FairDeDup and the SemDeDup maximum distance selection heuristic in Fig. 4, and show additional examples in the appendix.

4. Experiments

To assess the effect of deduplication on learned VLP models, we train CLIP-style models on variants of LAION-400M [58] and evaluate their performance on both standard and fairness-oriented benchmarks for zero-shot classification and text-image retrieval.

4.1. Models and Training

We train all models on LAION-400M [58] as a web-scale dataset representative of those typically used for large-scale vision-language pretraining. LAION-400M contains image-text pairs extracted from Common Crawl[1] filtered to

[1] https://commoncrawl.org/

(a) Maximum Distance Selection

(b) FairDeDup Selection

Figure 4. A random sampling of preserved samples from a cluster primarily composed of medical professionals after deduplication. FairDeDup improves selection diversity featuring increased variability in age, skin tone, and gender presentation.

have image-text CLIP similarity $\geq$0.3 without significant further curating. This makes LAION-400M an ideal test case for our setting as it is sufficiently large to train VLP models, captures bias from the internet, and is expected to contain semantically redundant samples. At the time of our data collection, only 375M image-text pairs from LAION-400M were still available for download.

We train CLIP-ViT-Base/16 [55] models from the OpenClip [30] implementation with vision transformer [18] base (ViT-B-16) as the image encoder and text transformer [67] as the text encoder. We perform distributed training over 80-120 A100 GPUs depending on the model with a global batch size of 33,820 image-caption pairs for 16 epochs regardless of dataset size. We use the AdamW [46] optimizer with linear warm up and cosine annealed learning rate schedule peaking at 5×10^{-4}. Additional hyperparameter details are provided in the appendix.

We evaluate CLIP training on three LAION-400M data settings for *performance* and *fairness*:

Baseline: LAION-400M. We train a CLIP model on the full LAION-400M dataset for a total of 183k steps as a control by which to evaluate baseline performance and fairness. A good model in the deduplicated setting should perform similarly to this model on common benchmarks without negatively impacting subgroup disparity and skew.

SemDeDup LAION-400M. For SemDeDup, we use a CLIP-ViT-Base/16 trained on WebImageText (WIT) [55] to produce image embedding which are partitioned into $50,000$ clusters using the FAISS [33] implementation of k-means and set the ϵ threshold for identifying duplicates within each cluster such that 50% of samples are pruned.

FairDeDup LAION-400M. We leverage the same WIT trained CLIP model for FairDeDup as SemDeDup. We consider 110 sensitive concepts capturing intersectional combinations of age, gender, skin tone, race and ethnicity and represent them using embeddings of 330 corresponding captions (three each with minor syntactic variation). We use the average across captions of the same concept as our prototypes. We enumerate all sensitive concepts and templates used to generate the text prototypes in the appendix. The selection step can be parallelized across CPUs up to the number of clusters produced by k-means. We find that selection in this setting on a 32 CPU machine takes one hour on average and that the overall time is dominated by the shared GPU parallelizable embedding and clustering steps.

4.2. Datasets and Metrics

We evaluate across three benchmarks to validate if models trained on deduplicated data are both *performant* and *fair*.

Zero-Shot Classification and Retrieval. We evaluate the performance of each model across 41 common zero-shot classification and retrieval datasets from Clip Benchmark [12] such as ImageNet [16], Flickr30k [71], and VTAB [73]. A model trained on deduplicated data should perform at least as well as a model trained in the full-data setting on these benchmarks.

Fair Zero-Shot Classification. The FACET [26] dataset contains expert reviewer annotations for 52 person related classes, gender presentation, skin tone, age, and other attributes, on a 32k image subset of Segement Anything 1 Billion (SA-1B) [36]. We perform zero-shot classification over the 52 person-classes by constructing a text prompt (*e.g.*, "*A photo of a* $\{class\}$") for each class and predicting the class used to construct the prompt with highest similarity to the image. Given a model f, sensitive attribute label l, person-class $\mathcal{C}$, and set of of images $\mathcal{I}_l^{\mathcal{C}}$ which captures class $\mathcal{C}$

featuring a person with label l, we measure the average and worst-class disparity in recall between subgroups of sensitive attributes where disparity is defined as:

$$disparity = recall(f(l_1, \mathcal{I}^C_{l_1}, \mathcal{C})) - recall(f(l_2, \mathcal{I}^C_{l_2}, \mathcal{C})). \quad (2)$$

Conceptually, a large magnitude disparity indicates that a model better predicts positive instances of person-class $\mathcal{C}$ for one of the two subgroups, while a disparity of zero indicates *equality of opportunity* between subgroups.

We evaluate subgroup disparity for average perceived gender expression by masculine *vs*. feminine presentation, lighter (1-4MST[2]) *vs*. darker (6-10MST) skin tone, and middle *vs*. younger and middle *vs*. older age for all person-classes which have at least 25 samples in both subgroups. Gustafson et al. [26] consider only the 21k images capturing a single person in their disparity analysis for simplicity and alignment between tasks (*e.g*., classification and visual grounding). To increase the sample size of our analysis, we consider each person in the dataset as a unique sample. We expand the bounding box for each person by 20% to capture context before extracting a square-padded image crop centered on the box, yielding 49,551 images.

Fair Image Retrieval. FairFace [34] annotates a balanced dataset of 108k cropped faces from YFCC-100M [66] by seven racial groups with additional annotations for perceived gender and age. Similar to [3, 14, 60], we measure the degree to which the top-k results of an image-text query differ over values of sensitive attributes in the 11k image validation set with respect to the desired proportion of those values with MaxSkew@1000 [24]. Given the top-k images τ^k_r returned by image-text query r, let the actual proportion of images returned by the query for a particular value $a_i \in A$ of sensitive attribute A be $P_{\tau^k_r, r, a_i} \in [0, 1]$ and the desired proportion be $P_{q,r,a_i} \in [0, 1]$, then the skew of value a_i is:

$$Skew_{a_i}@k(\tau_r) = \ln\left(\frac{P_{\tau^k_r, r, a_i}}{P_{q,r,a_i}}\right) \quad (3)$$

One limitation of Skew@k is that it is defined only for a single value of a sensitive attribute. To give a more holistic view across all values that a sensitive attribute may take on, we report the most skewed a_i with MaxSkew@k:

$$MaxSkew@k(\tau_r) = \max_{a_i \in A} Skew_{a_i}@k(\tau_r) \quad (4)$$

Conceptually, MaxSkew indicates the "*largest unfair advantage*" [24] provided to images with a particular value of the sensitive attribute for appearing in the the top-k results of the query. We choose the desired proportion of images to be the same as the true distribution of those images in the dataset. Under this condition, if the proportion of of a_i in the top-k results is the same as its distribution in the dataset, MaxSkew obtains an optimal result of 0 and achieves *demographic parity*. Following [3], we report average MaxSkew@1000 across 240 (un)favorable captions orthogonal to images in the dataset (*e.g*., "*A photo of a* {*smart*} *person*"), matching test attributes and prompts for race ($|A|$=7), gender ($|A|$=2), and age ($|A|$=3). Similar to Seth et al. [60], we bin age into larger groups: *younger* (0-19), *middle* (20-49), and *older* (50-70+) to reduce noise.

We additionally report MinSkew@k, which captures the "*worst disadvantage in representation*" for a subgroup, and the normalized discounted cumulative KL-divergence (NDKL), which captures the weighted average of Skew@k over all attribute values at varying settings of k. Intuitively, MinSkew captures the severity of the most negatively biased subgroup juxtaposed against the most positively biased captured by MaxSkew, and NDKL is a summary statistic over configurations of Skew@k. We refer readers to Geyik et al. [24] for the formulation of MinSkew and NDKL.

[2]Monk Skin Tone scale [51]

	Full Data (100%)	**SemDeDup** (50%)	**FairDeDup** (50%)
IN1K$_{acc@5}$	.899	.897 (–.002)	.897 (–.002)
INV2$_{acc@5}$	.845	.841 (–.004)	.837 (–.008)
C10$_{acc@5}$	.999	.998 (–.001)	.999 (–.000)
C100$_{acc@5}$	.934	.934 (–.000)	.939 (+.005)
Flickr$_{R@5}$	.873	.874 (+.001)	.871 (–.002)
COCO$_{R@5}$	.633	.632 (–.001)	.626 (–.007)

Table 1. Common zero-shot and text-image retrieval benchmarks for CLIP models on ImageNet1K [16], ImageNetV2 [56], CIFAR [39] (C10/C100), Flicker30k [71], and COCO Captions [43]. Higher (↑) is better in all cases. The difference in performance from the full-data setting is shown in green (red) when improved (reduced). Both deduplication strategies yield models that preserve the performance of models trained on the full data.

5. Results

Deduplication Preserves Aggregate Performance. In Tab. 1, we report Accuracy@5 for four common zero-shot image classification datasets: ImageNet1K [16], ImageNetV2 [56], CIFAR-10 and CIFAR-100 [39], and Recall@5 for two common image-text retrieval datasets: Flicker30k [71] and COCO Captions [43]. As expected, the performance drop from the full-data setting to deduplicated is marginal ($\leq$0.8%), indicating that performance is preserved after pruning 50% of the training data. The performance gap between the two deduplicated-data models is even smaller ($\leq$0.6%), and neither consistently performs more favorably across tasks. We refer readers to the appendix for results on additional datasets and metrics.

		Full (100%)	SemDeDup (50%)	FairDeDup (50%)	Diff. (FDD-SDD)
Gender Masc / Femn	Mean	.104	.113 (+ 9%)	**.109 (+ 5%)**	–.004
	Max	.303	.346 (+14%)	**.298 (– 2%)**	–.048
	Gap	.199	.233 (+17%)	**.189 (– 5%)**	–.053
Skin Tone Light / Dark	Mean	.100	.112 (+12%)	**.105 (+ 5%)**	–.007
	Max	.354	.342 (– 3%)	**.320 (–10%)**	–.022
	Gap	.254	.230 (– 9%)	**.215 (–15%)**	–.015
Age Mid / Yng	Mean	.063	**.059 (– 6%)**	.075 (+19%)	+.016
	Max	.268	.230 (–14%)	**.225 (–16%)**	–.005
	Gap	.205	.171 (–17%)	**.150 (–27%)**	–.022
Age Mid / Old	Mean	.098	.096 (– 2%)	**.087 (–11%)**	–.009
	Max	.252	.248 (– 2%)	**.153 (–39%)**	–.095
	Gap	.154	.152 (– 1%)	**.066 (–57%)**	–.008

Table 2. Absolute disparity (Eq. 2) in zero-shot classification performance on FACET [26] averaged across 52 person classes. Larger values indicate a greater performance gap between subgroups when predicting true positive samples of the same occupation. Lower (↓) is better for all metrics. Best deduplicated model in **bold**. The percent change in fairness outcomes from the full-data setting is shown in green (red) when improved (reduced).

SemDeDup Has Mixed Effects on Fairness. We show the result of our zero-shot image classification evaluation on FACET [26] in Tab. 2, studying subgroups across gender, skin tone, and age-based sensitive attributes. We find that SemDeDup yields mixed impacts. SemDeDup reinforces average and worst-class disparity across gender subgroups, exacerbates average disparity in skin tone while mitigating the worst-class, and surprisingly aids in reducing average and worst-class disparity across age groups.

In Tab. 3, we present the results of our text-image retrieval evaluation on FairFace [34], focusing on subgroups related to gender, race, and age. We again find that SemDeDup demonstrates mixed effects. SemDeDup reinforces gender skew across all metrics but mitigates skew towards the largest unfairly advantaged group (MaxSkew) while magnifying skew away from the worst disadvantaged group.

FairDeDup Improves Fairness Over SemDeDup. FairDeDup improves fairness outcomes over SemDeDup on FACET by mitigating, rather than exasperating, worst-class gender disparity while improving disparity outcomes in all cases except for age between middle-aged and young subgroups. With respect to SemDeDup, FairDeDup reduces the average over groups for mean disparity by .0001 (.0067 excluding `Age Mid/Yng`), worst-class by .0425 and gap by .0245. This result demonstrates that FairDeDup more closely achieves *equality of opportunity* than SemDeDup.

FairFace also shows evidence that FairDeDup improves fairness outcomes. While both methods increase gender skew, FairDeDup exhibits a milder skew across all summary metrics. For race, both methods mitigate the effects of the largest unfairly advantaged group (MaxSkew) compared to the baseline, while FairDeDup mitigates the magnitude of MaxSkew and reduces the skew against the worst disadvantaged class (MinSkew) compared to the baseline. Determining the best-performing method for age-based subgroups is inconclusive. Across gender and race groups, FairDeDup reduces MinSkew by .0555, MaxSkew by .0285 and NDKL by .0035. This results demonstrates that FairDeDup better achieves *demographic parity* than SemDeDup w.r.t. gender and race, even outperforming the full-data setting on race.

		Full (100%)	SemDeDup (50%)	FairDeDup (50%)	Diff. (FDD-SDD)
Gender	MinSkew	.159	.223 (+40%)	**.182 (+14%)**	–.041
	MaxSkew	.123	.153 (+24%)	**.125 (+ 2%)**	–.028
	NDKL	.010	.015 (+50%)	**.012 (+20%)**	–.003
Race	MinSkew	.545	.583 (+ 7%)	**.513 (– 6%)**	–.070
	MaxSkew	.432	.401 (– 7%)	**.372 (–14%)**	–.029
	NDKL	.035	.034 (– 3%)	**.030 (–14%)**	–.004
Age	MinSkew	.618	.702 (+14%)	**.647 (+ 5%)**	–.055
	MaxSkew	.241	**.224 (– 7%)**	.296 (+23%)	+.072
	NDKL	.023	**.022 (– 4%)**	.028 (+22%)	+.006

Table 3. Skew evaluation on FairFace [34] averaged over 240 text-image query templates. As MinSkew is a negative metric optimal at its upper bound of zero, we report its absolute value for readability so that lower (↓) is better for all metrics. Best deduplicated model in **bold**. The percent change in fairness outcomes from the full-data setting is shown in green (red) when improved (reduced).

6. Discussion

Below we discuss observations when pruning smaller-scale annotated data, potential FairDeDup variants for varied concept prototypes, and limitations of our approach.

Evaluation on Demographically Annotated Data. In this paper, we have shown on large-scale real model training that FairDeDup achieves results on-par with SemDeDup on standard benchmarks, while demonstrating improved fairness outcomes. We believe that is the clearest signal about the applicability of FairDeDup in real-world usage. However, we would also like to directly demonstrate that FairDeDup does indeed select more diverse data representations compared to SemDeDup. To do so, we consider deduplicating the FACET [26] images described in Sec. 4. We perform k-means clustering (k=50) on the images with ten different random seeds and apply both deduplication methods to each. In Tab. 4, we report the percent of the post-pruning dataset labeled as non-majority classes for gender (*feminine*, *non-binary*, *other*), skin tone (MST>4, *other*), and age (*younger*, *older*, *other*), averaged across the ten

	Full Data (100%)	SemDeDup (50%)	FairDeDup (50%)
Gender	32.92%	31.91%	**32.29%**
Skin Tone	51.28%	50.46%	**51.06%**
Age	44.74%	43.62%	**44.06%**

Table 4. Data mass allocated to minority classes in FACET [26] after deduplication averaged over ten random seeds. We consider minority classes by gender, skin tone, and age. The difference between means of SemDeDup and FairDeDup across trials is significant at $>99.9\%$ confidence (n=10) for all groups according to a paired t-test. In all cases, we observe that FairDeDup helps recover mass reallocated to the majority class by SemDeDup.

trials. This analysis indicates that 1) SemDeDup does indeed reduce the frequency of the least well represented subgroups and 2) that FairDeDup mitigates this effect. The difference between means across trials of SemDeDup and FairDeDup is statistically significant at $\geq 99.9\%$ confidence (n=10) across all groups according to a paired t-test.

Variants and Applications of the FairDedup Algorithm. In our experiments, we use text-based prototypes to guide FairDeDup towards balancing representation of sensitive concepts. However, the exact specification of these prototypes is flexible to other subjects (*e.g.*, non-person related) and modalities (*e.g.*, image-based concepts). FairDeDup can be trivially modified to consider any prototype for which the embedding model can output similarity to individual images, such as sets of semantically aligned images (e.g., based on image type, photographs, illustrations, infographics, etc), or a combination of image and text prototypes. Similarly, FairDedup can be used to boost underrepresented samples from arbitrary sets such as object entities [70], or other forms of semantic organization.

6.1. Limitations

Clustering Restrictions on Selection. While clustering allows deduplication algorithms to scale to hundreds of millions of samples, it also limits the availability of lower-represented samples for balancing sensitive concept representation. Take, for example, a data subset capturing photos of dancers. If the clustering algorithm creates two "dancer" clusters, bifurcating across binary gender presentation, then FairDeDup will be unable to perform significant gender balancing due to the independent processing of each cluster. We note that the resulting balance will be based on a combination of the underlying number of "dancer" photos in the dataset and the rate of duplication within both groups. If the two clusters are approximately equal sized with equal frequency of semantic duplicates, the independent deduplication of both clusters is equivalent in representation to a joint deduplication with respect to the bifurcated attribute. We display demonstrative clusters in the appendix.

Bias Transfer From the Embedding Model. By deduplicating based on model embeddings, we subject the selection of samples to the biases of the embedding model. The majority of sensitive concepts we select are social constructs based in gender and race, and are not identifiable by anyone other than the photographed individual. We therefore expect sensitive concept representation to be based upon the predominate social norms they capture, rather than necessarily true identities of individuals. Nonetheless, we assert that a deduplication method which maintains the bias of the full-data setting is a favorable start to one that magnifies it.

Demographic Representation in Fairness Datasets. Most contemporary fairness datasets lack annotations from the individuals they represent. Consequently, for nonstationary socially constructed attributes such as gender, race, and perceived *young/oldness*, the captured data relies solely on annotators' subjective understanding. Additionally, these datasets often limit gender representation to a binary perspective (occasionally including a small "other" category) [17], a necessary operationalization for scale that is not inclusive of bias characterization for diverse gender identities. We also note that fairness datasets cover a limited number of directions under which a model may express bias, excluding disability, national origin, and other sensitive attributes. Our analysis, therefore, only examines fairness outcomes with respect to contemporary and subjective evaluation of these limited available demographic attributes.

7. Conclusion

In this paper, we study the fairness outcomes resulting from training large-scale vision-language models on semantically deduplicated web-scale data, using LAION-400M and SemDeDup as a representative dataset and deduplication algorithm pairing. We find that deduplication has consistently harmful effects on gender-based bias and mixed effects on skin tone/race- and age-based biases across zero-shot classification and text-image retrieval tasks. To improve fairness outcomes, we propose FairDeDup, a simple and efficient fairness-aware modification of the sample selection heuristic in SemDeDup which boosts the representation of user-defined sensitive concepts in the post-deduplication data distribution. Our experiments show that FairDeDup preserves the performance of the full-data setting on standard metrics for common image-text datasets, has more favorable fairness outcomes than SemDeDup across all cases for gender- and skin tone/race-based biases, and outperforms the baseline full-data setting in several instances. We hope for FairDeDup to provide a simple and tractable baseline for future work in fairness-aware deduplication.

References

[1] Amro Abbas, Kushal Tirumala, Dániel Simig, Surya Ganguli, and Ari S. Morcos. Semdedup: Data-efficient learning at web-scale through semantic deduplication. *arXiv preprint arXiv:2303.09540*, 2023. 1, 2, 3

[2] Sandhini Agarwal, Gretchen Krueger, Jack Clark, Alec Radford, Jong Wook Kim, and Miles Brundage. Evaluating clip: towards characterization of broader capabilities and downstream implications. *arXiv preprint arXiv:2108.02818*, 2021. 2

[3] Hugo Berg, Siobhan Hall, Yash Bhalgat, Hannah Kirk, Aleksandar Shtedritski, and Max Bain. A prompt array keeps the bias away: Debiasing vision-language models with adversarial learning. In *Proceedings of the 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 12th International Joint Conference on Natural Language Processing*, pages 806–822, 2022. 2, 6

[4] Joseph R. Biden and The White House. Executive order on the safe, secure, and trustworthy development and use of artificial intelligence, 2023. Available: https://www.whitehouse.gov/briefing-room/presidential-actions/2023/10/30/executive-order-on-the-safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence/. 3

[5] Abeba Birhane, Vinay Uday Prabhu, and Emmanuel Kahembwe. Multimodal datasets: misogyny, pornography, and malignant stereotypes. *arXiv:2110.01963 [cs]*, 2021. 2

[6] Abeba Birhane, Vinay Prabhu, Sang Han, Vishnu Naresh Boddeti, and Alexandra Sasha Luccioni. Into the laions den: Investigating hate in multimodal datasets. *Neural Information Processing Systems*, 2023. 2

[7] Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna Wallach. Language (technology) is power: A critical survey of "bias" in NLP. In *Association for Computational Linguistics*, pages 5454–5476, 2020. 2

[8] Tolga Bolukbasi, Kai-Wei Chang, James Y. Zou, Venkatesh Saligrama, and Adam Tauman Kalai. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Neural Information Processing Systems*, 2016. 2

[9] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *ACM Conference on Fairness Acountability and Transparency*, 2018. 2

[10] Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186, 2017. 2

[11] Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Alexander Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, Alexander Kolesnikov, Joan Puigcerver, Nan Ding, Keran Rong, Hassan Akbari, Gaurav Mishra, Linting Xue, Ashish Thapliyal, James Bradbury, Weicheng Kuo, Mojtaba Seyedhosseini, Chao Jia, Burcu Karagol Ayan, Carlos Riquelme, Andreas Steiner, Anelia Angelova, Xiaohua Zhai, Neil Houlsby, and Radu Soricut. Pali: A jointly-scaled multilingual language-image model. In *International Conference on Learning Representations*, 2023. 2

[12] Mehdi Cherti and Romain Beaumont. Clip benchmark, 2023. GitHub repository https://github.com/LAION-AI/CLIP_benchmark. 1, 2, 5

[13] Mehdi Cherti, Romain Beaumont, Ross Wightman, Mitchell Wortsman, Gabriel Ilharco, Cade Gordon, Christoph Schuhmann, Ludwig Schmidt, and Jenia Jitsev. Reproducible scaling laws for contrastive language-image learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 2818–2829, 2023. 1

[14] Ching-Yao Chuang, Varun Jampani, Yuanzhen Li, Antonio Torralba, and Stefanie Jegelka. Debiasing vision-language models via biased prompts. *arXiv preprint arXiv:2302.00070*, 2023. 2, 6

[15] U.S. Equal Employment Oppourtunity Comission. Title vii of the civil rights act of 1964, 1964. Available: https://www.eeoc.gov/statutes/title-vii-civil-rights-act-1964. 3

[16] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition*, pages 248–255, 2009. 1, 3, 5, 6

[17] Hannah Devinney, Jenny Björklund, and Henrik Björklund. Theories of "gender" in nlp bias research. In *ACM Conference on Fairness, Accountability, and Transparency*, pages 2083–2102, 2022. 8

[18] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. 5

[19] Conor Dougherty. Google photos mistakenly labels black people 'gorillas'. *New York Times*, 2015. 2

[20] Ling Du, Anthony T.S. Ho, and Runmin Cong. Perceptual hashing for image authentication: A survey. *Signal Processing: Image Communication*, page 115713, 2020. 3

[21] Jade S. Franklin, Karan Bhanot, Mohamed Ghalwash, Kristin P. Bennett, Jamie McCusker, and Deborah L. McGuinness. An ontology for fairness metrics. In *AAAI/ACM Conference on AI, Ethics, and Society*, page 265–275, 2022. 3

[22] Samir Yitzhak Gadre, Gabriel Ilharco, Alex Fang, Jonathan Hayase, Georgios Smyrnis, Thao Nguyen, Ryan Marten, Mitchell Wortsman, Dhruba Ghosh, Jieyu Zhang, Eyal Orgad, Rahim Entezari, Giannis Daras, Sarah Pratt, Vivek Ramanujan, Yonatan Bitton, Kalyani Marathe, Stephen Mussmann, Richard Vencu, Mehdi Cherti, Ranjay Krishna, Pang Wei Koh, Olga Saukh, Alexander Ratner, Shuran Song, Hannaneh Hajishirzi, Ali Farhadi, Romain Beaumont, Sewoong Oh, Alex Dimakis, Jenia Jitsev, Yair Carmon, Vaishaal Shankar, and Ludwig Schmidt. Datacomp: In search of the next generation of multimodal datasets. *arXiv preprint arXiv:2304.14108*, 2023. 3

[23] Noa Garcia, Yusuke Hirota, Yankun Wu, and Yuta Nakashima. Uncurated image-text datasets: Shedding light on demographic bias. In *Computer Vision and Pattern Recognition*, pages 6957–6966, 2023. 2

[24] Sahin Cem Geyik, Stuart Ambler, and Krishnaram Kenthapadi. Fairness-aware ranking in search & recommendation systems with application to linkedin talent search. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, page 2221–2231, 2019. 6

[25] Chengcheng Guo, Bo Zhao, and Yasnbing Bai. Deepcore: A comprehensive library for coreset selection in deep learning. In *International Conference on Database and Expert Systems Applications*, pages 181–195. Springer, 2022. 2

[26] Laura Gustafson, Chloe Rolland, Nikhila Ravi, Quentin Duval, Aaron Adcock, Cheng-Yang Fu, Melissa Hall, and Candace Ross. Facet: Fairness in computer vision evaluation benchmark. In *International Conference on Computer Vision*, pages 20370–20382, 2023. 1, 2, 5, 6, 7, 8

[27] Lisa Anne Hendricks, Kaylee Burns, Kate Saenko, Trevor Darrell, and Anna Rohrbach. Women also snowboard: Overcoming bias in captioning models. In *Neural Information Processing Systems*, 2016. 2

[28] Yusuke Hirota, Yuta Nakashima, and Noa Garcia. Quantifying societal bias amplification in image captioning. In *Computer Vision and Pattern Recognition*, pages 13450–13459, 2022. 2

[29] IEEE. Ieee code of ethics, 2020. Available: https://www.ieee.org/about/corporate/governance/p7-8.html. 3

[30] 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, 2021. Zendo https://doi.org/10.5281/zenodo.5143773. 5

[31] Sepehr Janghorbani and Gerard De Melo. Multimodal bias: Introducing a framework for stereotypical bias assessment beyond gender and race in vision language models. *European Chapter of the Association for Computational Linguistics*, 2023. 2

[32] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916, 2021. 2, 4

[33] Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547, 2019. 5

[34] Kimmo Kärkkäinen and Jungseock Joo. Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *Winter Conference on Applications of Computer Vision*, pages 1548–1558, 2021. 2, 6, 7

[35] Justin Kerr, Chung Min Kim, Ken Goldberg, Angjoo Kanazawa, and Matthew Tancik. Lerf: Language embedded radiance fields. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 19729–19739, 2023. 1

[36] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, and Ross Girshick. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023. 5

[37] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C Berg, Wan-Yen Lo, et al. Segment anything. *arXiv preprint arXiv:2304.02643*, 2023. 1

[38] Svetlana Kiritchenko and Saif M. Mohammad. Examining gender and race bias in two hundred sentiment analysis systems. *Conference on Lexical and Computational Semantics*, 2018. 2

[39] Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images, 2009. 6

[40] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. Align before fuse: Vision and language representation learning with momentum distillation. *Neural Information Processing Systems*, pages 9694–9705, 2021. 2

[41] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*, pages 12888–12900, 2022. 4

[42] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. *arXiv preprint arXiv:2301.12597*, 2023. 1

[43] 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, 2014. 1, 6

[44] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. Improved baselines with visual instruction tuning. *arXiv preprint arXiv:2310.03744*, 2023. 1

[45] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, 2015. 2

[46] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. 5

[47] Ryan Mac. Facebook apologizes after a.i. puts 'primates' label on video of black men. *New York Times*, 2021. 2

[48] Adyasha Maharana, Prateek Yadav, and Mohit Bansal. D2 pruning: Message passing for balancing diversity & difficulty in data pruning. *arXiv preprint arXiv:2310.07931*, 2023. 2, 3

[49] Chandler May, Alex Wang, Shikha Bordia, Samuel R Bowman, and Rachel Rudinger. On measuring social biases in sentence encoders. *North American Chapter of the Association for Computational Linguistics*, 2019. 2

[50] Baharan Mirzasoleiman, Jeff Bilmes, and Jure Leskovec. Coresets for data-efficient training of machine learning models. In *International Conference on Machine Learning*, pages 6950–6960, 2020. 2

[51] Ellis Monk. Monk skin tone scale, 2019. 6

[52] Safiya Noble. *Algorithms of Oppression: How Search Engines Reinforce Racism*. New York University Press, Cambridge MA, 2018. 2

[53] Council of European Union. Directive Proposal (COM(2008)462), 2008. Available: https://eur-lex.europa.eu/legal-content/en/TXT/?uri=CELEX%3A52008PC0426. 3

[54] Jeff M. Phillips. Coresets and sketches. *arXiv preprint arXiv:1601.00617*, 2016. 3

[55] Alec Radford, Jong Wook Kim, Chris Hallacy, A. Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, 2021. 1, 2, 3, 4, 5

[56] Benjamin Recht, Rebecca Roelofs, Ludwig Schmidt, and Vaishaal Shankar. Do ImageNet classifiers generalize to ImageNet? In *International Conference on Machine Learning*, pages 5389–5400, 2019. 6

[57] Candace Ross, Boris Katz, and Andrei Barbu. Measuring social biases in grounded vision and language embeddings. In *North American Chapter of the Association for Computational Linguistics*, pages 998–1008, 2021. 2

[58] 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. *NeurIPS Workshop on Data Centric AI*, 2021. 1, 2, 4

[59] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, et al. Laion-5b: An open large-scale dataset for training next generation image-text models. *Neural Information Processing Systems*, 35:25278–25294, 2022. 1

[60] Ashish Seth, Mayur Hemani, and Chirag Agarwal. Dear: Debiasing vision-language models with additive residuals. In *Computer Vision and Pattern Recognition*, pages 6820–6829, 2023. 2, 6

[61] Nur Muhammad Mahi Shafiullah, Chris Paxton, Lerrel Pinto, Soumith Chintala, and Arthur Szlam. Clip-fields: Weakly supervised semantic fields for robotic memory. *arXiv preprint arXiv:2210.05663*, 2022. 1

[62] Mohit Shridhar, Lucas Manuelli, and Dieter Fox. Cliport: What and where pathways for robotic manipulation. In *Conference on Robot Learning*, pages 894–906. PMLR, 2022. 1

[63] Ben Sorscher, Robert Geirhos, Shashank Shekhar, Surya Ganguli, and Ari Morcos. Beyond neural scaling laws: beating power law scaling via data pruning. In *Neural Information Processing Systems*, pages 19523–19536, 2022. 2, 3, 4

[64] Tejas Srinivasan and Yonatan Bisk. Worst of both worlds: Biases compound in pre-trained vision-and-language models. In *Proceedings of the 4th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 77–85, 2022. 2

[65] Tony Sun, Andrew Gaut, Shirlyn Tang, Yuxin Huang, Mai ElSherief, Jieyu Zhao, Diba Mirza, Elizabeth Belding, Kai-Wei Chang, and William Yang Wang. Mitigating gender bias in natural language processing: Literature review. In *Association for Computational Linguistics*, pages 1630–1640, 2019. 2

[66] 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*, pages 64–73, 2016. 6

[67] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Neural Information Processing Systems*, 30, 2017. 5

[68] Angelina Wang, Solon Barocas, Kristen Laird, and Hanna Wallach. Measuring representational harms in image captioning. In *ACM Conference on Fairness, Accountability, and Transparency*, pages 324–335, 2022. 2

[69] Ted Xiao, Harris Chan, Pierre Sermanet, Ayzaan Wahid, Anthony Brohan, Karol Hausman, Sergey Levine, and Jonathan Tompson. Skill acquisition by instruction augmentation on offline datasets. In *NeurIPS 2022 Foundation Models for Decision Making Workshop*, 2022. 1

[70] Hu Xu, Saining Xie, Xiaoqing Ellen Tan, Po-Yao Huang, Russell Howes, Vasu Sharma, Shang-Wen Li, Gargi Ghosh, Luke Zettlemoyer, and Christoph Feichtenhofer. Demystifying clip data. *arXiv preprint arXiv:2309.16671*, 2023. 8

[71] 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*, pages 67–78, 2014. 1, 5, 6

[72] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. Coca: Contrastive captioners are image-text foundation models. *Transactions on Machine Learning Research*, 2022. 4

[73] Xiaohua Zhai, Joan Puigcerver, Alexander Kolesnikov, Pierre Ruyssen, Carlos Riquelme, Mario Lucic, Josip Djolonga, André Susano Pinto, Maxim Neumann, Alexey Dosovitskiy, Lucas Beyer, Olivier Bachem, Michael Tschannen, Marcin Michalski, Olivier Bousquet, Sylvain Gelly, and Neil Houlsby. The visual task adaptation benchmark. *arXiv preprint arXiv:1910.04867*, 2019. 1, 5

[74] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. Vinvl: Revisiting visual representations in vision-language models. In *Computer Vision and Pattern Recognition*, pages 5579–5588, 2021. 4

[75] Yi Zhang, Junyang Wang, and Jitao Sang. Counterfactually measuring and eliminating social bias in vision-language pre-training models. In *ACM International Conference on Multimedia*, pages 4996–5004, 2022. 2

[76] Dora Zhao, Angelina Wang, and Olga Russakovsky. Understanding and evaluating racial biases in image captioning. In *International Conference on Computer Vision*, pages 14830–14840, 2021. 2

[77] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Conference on Empirical Methods in Natural Language Processing*, 2017. 2

[78] Kankan Zhou, Yibin LAI, and Jing Jiang. Vlstereoset: A study of stereotypical bias in pre-trained vision-language models. In *Association for Computational Linguistics*, 2022. 2

[79] Xueyan Zou, Jianwei Yang, Hao Zhang, Feng Li, Linjie Li, Jianfeng Gao, and Yong Jae Lee. Segment everything everywhere all at once. *arXiv preprint arXiv:2304.06718*, 2023. 1