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

Project Page

[1]Oregon State University
[2]Adobe Research

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

CVPR SEATTLE, WA JUNE 17-21, 2024

TL;DR: FairDeDup **mitigates bias** in data deduplication by preserving human-defined dimensions of diversity while retaining the ability to remove redundant samples for **faster model training.**

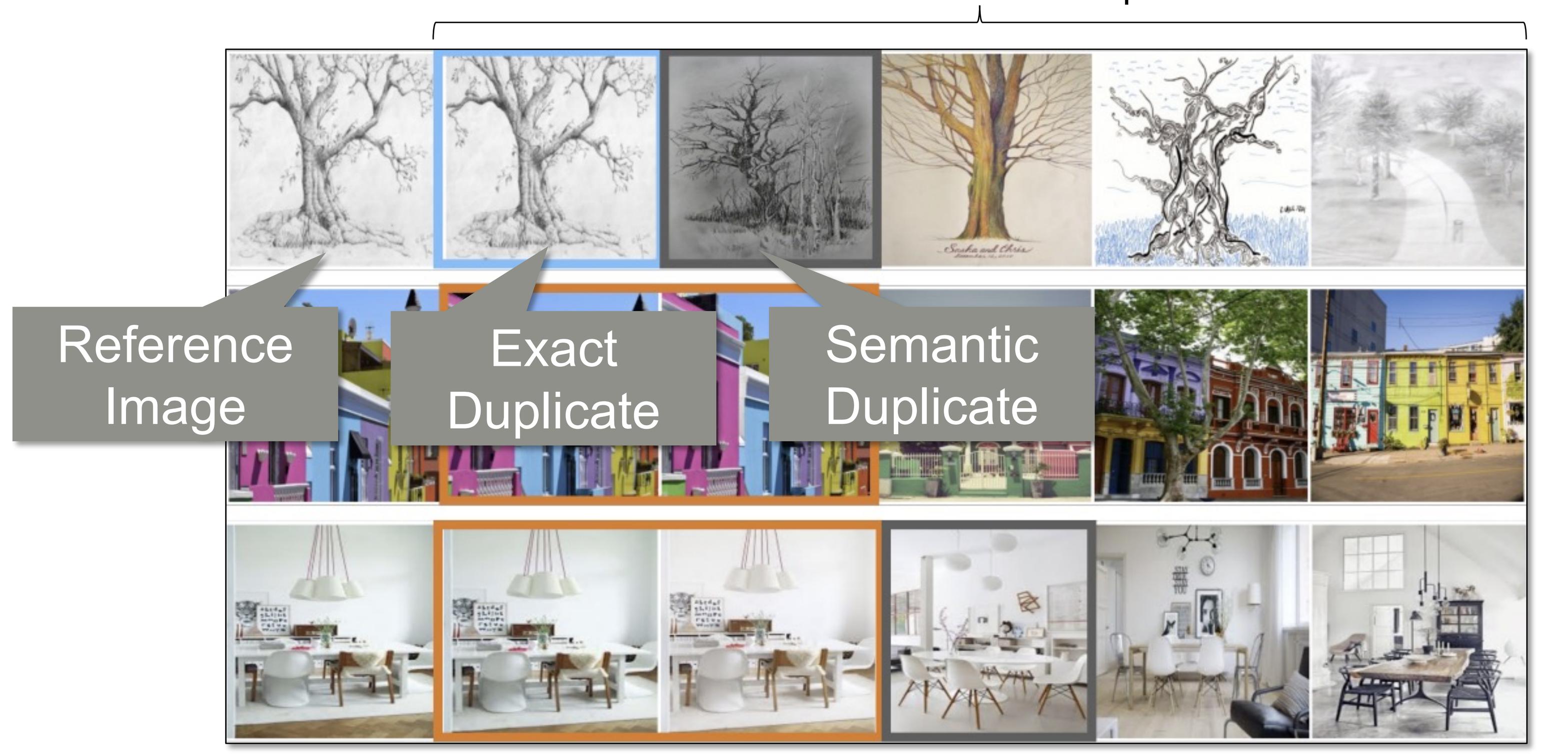
Observation: Web-Scale VL Models Are Expensive

LAION used 824 GPUs for 11 days to train one large model (CLIP H/14-4B). Est. On-Demand EC2 GPU cost alone for training is

$870K over 11 Days

Mitigation: Remove Web Training Data Duplicates

Removing duplicated data improves training efficiency by preventing redundant inputs. SemDeDup (SDD) [1] is a SOTA method that reduces training cost and time by half with minimal impact on model accuracy.



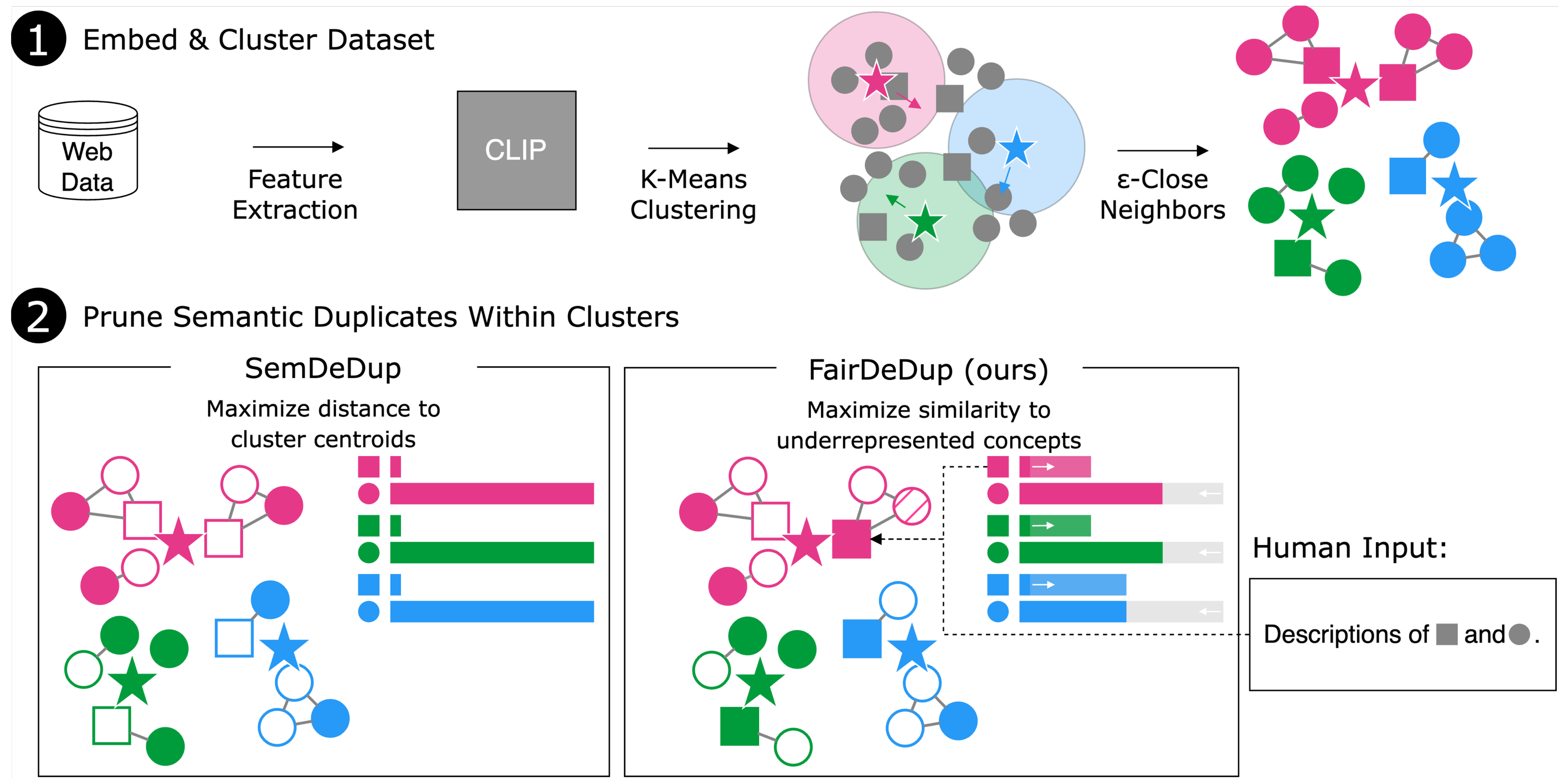
Problem: Deduplication Can Reinforce VL Bias

SDD can exasperate fairness disparities, reducing the range of genders, races, or ages seen in preserved images depicting certain occupations.

A slice of data from SDD showing lack of intersectional diversity across gender, skin tone, and age. Random samples selected from a large hand-picked cluster.

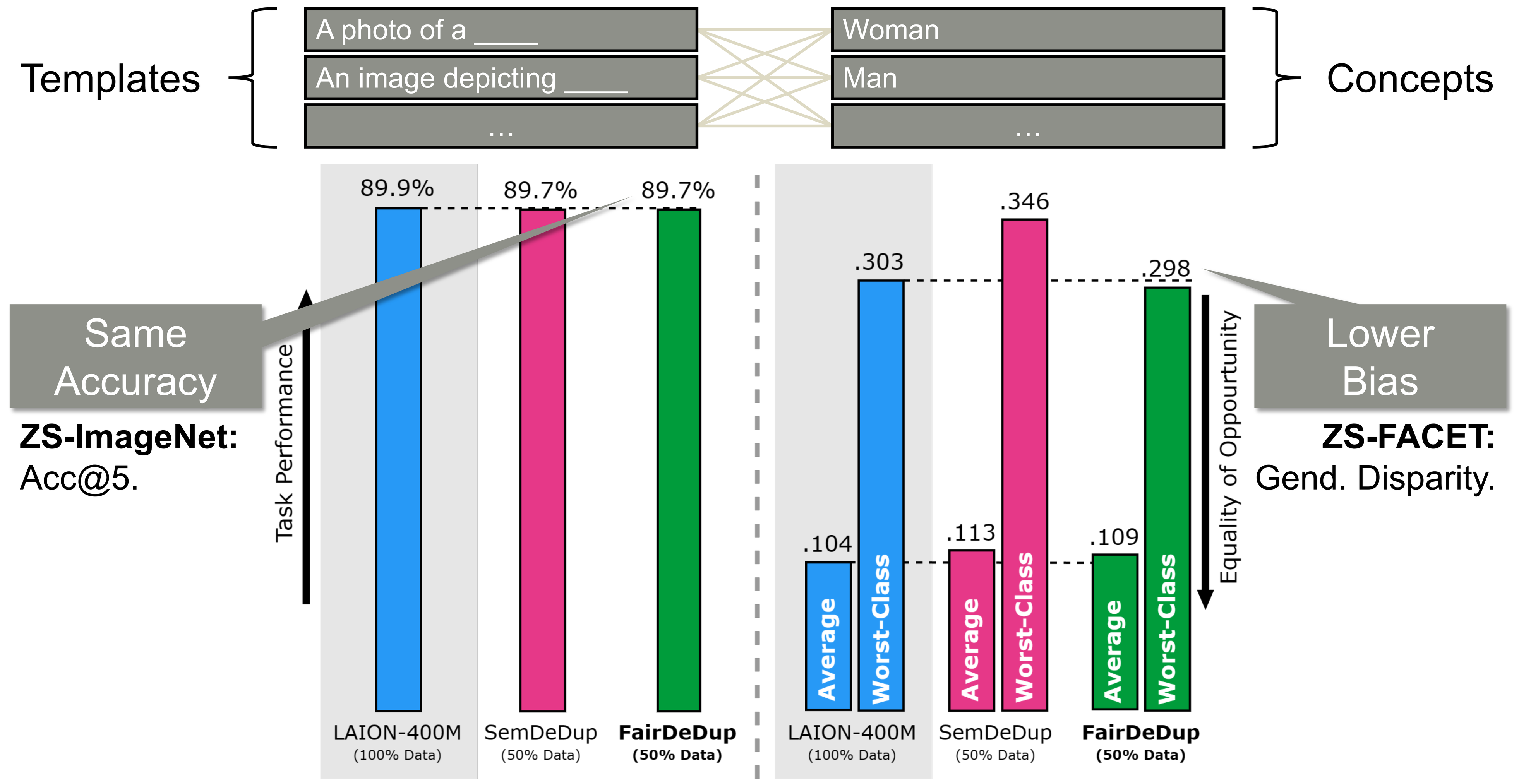
Our Solution: FairDeDup (FDD)

FDD preserves data diversity along human-defined semantic concepts specified in natural language. Concepts are incorporated into the sample preservation heuristic applied to neighborhoods of duplicated data.



Example: Mitigate Gender Bias & Preserve Accuracy

A human writes captions describing different genders. FDD uses these captions to improve diversity in samples preserved during deduplication.

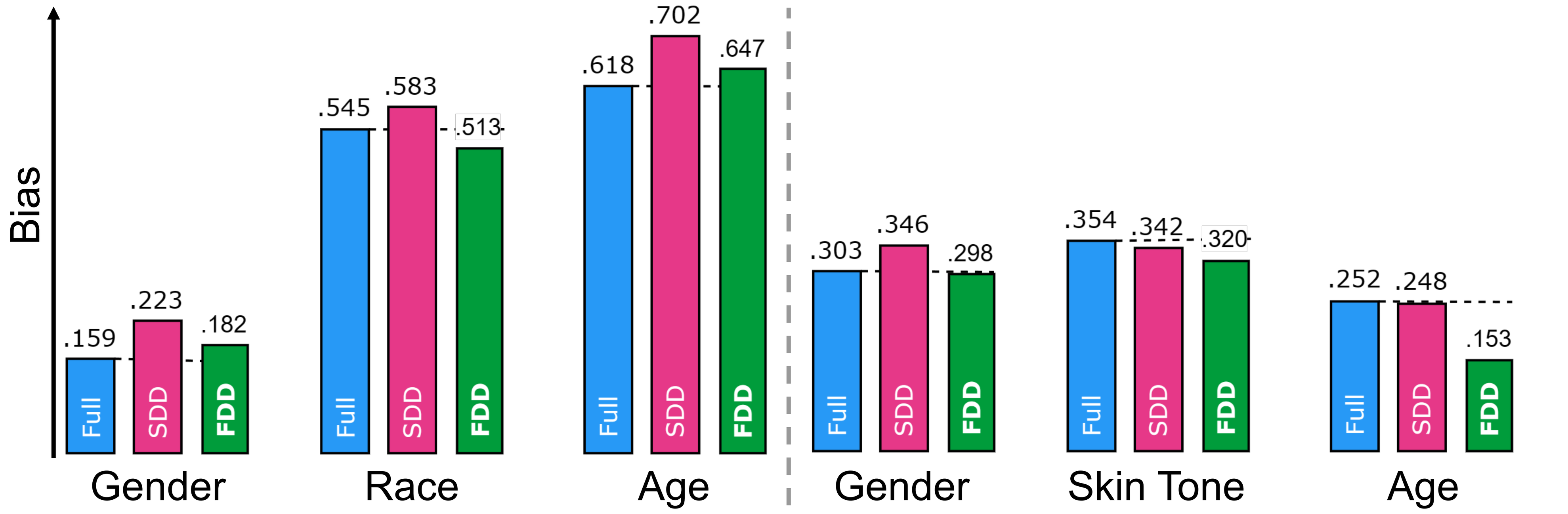


Result: FDD Preserves The Data Distribution

- We measure the %-data allocated to all non-majority classes across repeated runs on a smaller dataset with annotated demographics [2].
- Strong evidence ($p < .001$) from a paired t-test ($n = 10$) suggests a difference in %-data allocated between FDD and SDD.
- ~0.5% difference in %-data, equivalent to 2M samples in LAION-400M.

Result: FDD Reduces Downstream Bias

FDD improves fairness outcomes over SDD in nearly every case. The best performing deduplicated model is in **bold**. Lower is better.



FairFace/MinSkew: Bias against most disadvantaged group in a zero-shot text-based image query task.

FACET/Disparity: Worst-case bias against a group in a zero-shot text-based image classification task.

[1] Abbas et al. SemDeDup: Data-Efficient Learning at Web-Scale Through Semantic Deduplication. arXiv preprint, 2023.
[2] Gustafson et al. FACET: Fairness in Computer Vision Evaluation Benchmark. ICCV 2023.