Measuring Text-Image Retrieval Fairness with Synthetic Data

Lluis Gomez
lgomez@cvc.uab.es
Universitat Autònoma de Barcelona, Computer Vision Center.
Barcelona, Spain

Figure 1: Samples of the SISPI dataset for social bias assessment in cross-modal text-image retrieval models. We show 8 relevant images for two text queries: "A photo of a CEO" (top row) and "A photo of a nurse" (bottom row). Images shown for each query are generated with Stable Diffusion XL [62] using the same initial seed, so that the joint distribution of unprotected attributes (background, lighting, pose, etc.) is roughly equal across protected attribute (ethnicity and gender) groups.

Abstract

In this paper, we study social bias in cross-modal text-image retrieval systems, focusing on the interaction between textual queries and image responses. Despite the significant advancements in cross-modal retrieval models, the potential for social bias in their responses remains a pressing concern, necessitating a comprehensive framework for assessment and mitigation. We introduce a novel framework for evaluating social bias in cross-modal retrieval systems, leveraging a new dataset and appropriate metrics specifically designed for this purpose. Our dataset, Social Inclusive Synthetic Professionals Images (SISPI), comprises 49K images generated using state-of-the-art text-to-image models, ensuring a balanced representation of demographic groups across various professional roles. We use this dataset to conduct an extensive analysis of social bias (gender and ethnic) in state of the art cross-modal retrieval deep models, including CLIP, ALIGN, BLIP, FLAVA, COCA, and many others. Using diversity metrics, grounded in the distribution of different demographic groups' images in the retrieval rankings, we provide a quantitative measure of fairness, facilitating a detailed analysis of models' behavior. Our work sheds light on biases present in current cross-modal retrieval systems and emphasizes the importance of training data curation, providing a foundation for future research and development towards more equitable and unbiased models. The dataset and code of our framework is publicly available at https://sispi-benchmark.github.io/.

SIGIR '25, Padua, Italy

ACM ISBN 979-8-4007-1592-1/2025/07
https://doi.org/10.1145/3726302.3730030

CCS Concepts

• **Information systems** → **Information retrieval**; *Image search*; **Test collections**; • **Social and professional topics** → *Race and ethnicity*; **Gender**.

Keywords

Text-Image Retrieval, Fairness, Synthetic Dataset, CLIP

ACM Reference Format:
Lluis Gomez. 2025. Measuring Text-Image Retrieval Fairness with Synthetic Data. In *Proceedings of the 48th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR '25), July 13–18, 2025, Padua, Italy*. ACM, New York, NY, USA, 12 pages. https://doi.org/10.1145/3726302.3730030

1 Introduction

Cross-modal retrieval systems enable the retrieval of information across different modalities. The text-image retrieval task involves using a text query to find matching images, and vice versa. Such systems have gained significant momentum in recent years as a response to the challenge of searching for data in increasingly larger multi-modal databases.

State-of-the-art deep learning cross-modal retrieval models [1, 45–47, 63] achieve remarkable performance in standard benchmarks, such as MS-COCO [52] and Flickr30k [84]. As reliance on these systems increases in a myriad of widespread applications, it is crucial to identify and quantify potential social biases – such as incorrectly and consistently portraying certain activities or roles as predominantly associated with a particular ethnicity and/or gender – reinforcing non-inclusive stereotypes. Understanding and addressing these biases is essential to ensure these systems are fair, unbiased, and representative of the diverse users they serve.

Efforts like implementing fairness metrics, creating new datasets, and conducting diversity audits are steps towards mitigating these biases.

Although social biases are being increasingly studied in different areas of machine learning [4, 6, 11, 12, 21, 28, 32, 67, 76, 78] and information retrieval [2, 16, 23, 42, 58, 60, 61, 86], there is lack of a standard procedure to measure them systematically in cross-modal image-text retrieval models. One of the main challenges in this regard is the difficulty of obtaining an adequate large collection of diverse image-text pairs. Manually crafting a perfectly balanced dataset across various demographic groups, roles, and activities is a complicated task. Furthermore, even if such a dataset could be assembled, it could inadvertently contain visual artifacts – such as variations in background, human pose, clothes' color, etc. – that might affect the models' performance for different protected demographic groups [57].

In this paper, we propose a novel solution to address the aforementioned challenges by using synthetically generated data to measure social bias in cross-modal retrieval models. We take inspiration from related work that uses synthetic faces to benchmark face analysis [3] and recognition [50] systems. We build on their ideas, but address the specificities of cross-modal text-image retrieval models. By leveraging state of the art text-to-image generative models [62], we create a realistic, diverse, and balanced synthetic dataset of text-image pairs that circumvent the limitations of manual data collection methods. This approach allows us to precisely control not only protected attributes (ethnicity and gender) but also the unprotected attributes such as background settings, human poses, and object contexts.

The contributions of this paper include: (1) Introducing a framework to assess gender and ethnicity bias in cross-modal text-image retrieval, featuring a synthetic dataset and a well-established fairness metric; (2) Conducting an empirical analysis of state-of-the-art models to evaluate social bias and its relation to training datasets; and (3) Exploring the trade-offs between retrieval performance and fairness, setting the stage for future work on mitigating multiple social biases.

2 Related Work

Image-Text Retrieval. Image-text retrieval models align visual and textual representations within a shared subspace, enabling similarity calculations via simple distance metrics. Common approaches employ metric learning losses to achieve this alignment [24, 44, 48, 53, 77].

Recently, vision-language pre-training methods [19, 54, 55, 75, 88] have advanced cross-modal representation learning by pretraining on large and diverse datasets. This approach has not only improved image-text retrieval but also enhanced performance across various downstream tasks, such as visual question answering, image captioning, zero-shot classification, etc. Models like CLIP [63] and ALIGN [37] pushed the boundaries of this paradigm by scaling up vision-language representation learning with contrastive objectives on millions of (noisy) image-text pairs sourced from the Internet.

Social bias in cross-modal retrieval. Despite recent advancements in cross-modal image-text retrieval, a standard framework for systematically assessing social bias in these models is still lacking. This is particularly concerning as many of these models are trained on large, uncurated datasets known to contain non-inclusive biases and other problematic characteristics [5, 7–9, 27].

While most state-of-the-art vision-language pretraining methods acknowledge that their models are trained on noisy and biased datasets, few have rigorously assessed the impact of these biases on model behavior [7]. Radford *et al.* [63] explored potential social biases in the CLIP model within a zero-shot image classification context but did not extend this analysis to cross-modal retrieval. Similarly, the FACET (FAirness in Computer Vision EvaluaTion) benchmark [31] focuses on image classification, object detection, and segmentation, without addressing cross-modal retrieval. FACET includes 32k images from the Segment Anything 1 Billion (SA-1B) dataset [41], annotated with demographic labels by experts. In contrast, the PHASE (Perceived Human Annotations for Social Evaluation)[27] provides demographic annotations for a 19K subset of the Conceptual Captions (CC) dataset[71], offering a more tailored approach for evaluating text-image retrieval models. These annotations, combined with the original CC dataset, enable assessment of model performance variations across different demographic groups in cross-modal retrieval scenarios.

While FACET, PHASE, and similar efforts [70, 89, 90] are important steps towards fairness evaluation, they also have limitations due to their origin from existing real datasets. Besides the inherent costs of manual annotations, one major issue is the demographic imbalance. For instance, in FACET, the representation skews heavily towards more stereotypically maleness versus femaleness annotations (72% vs. 26%); and PHASE overrepresents White individuals compared to Middle Eastern or Southeast Asian descent (2,231 vs. 16 images). Another critical aspect is the challenge of estimating the causal effect of protected attributes on algorithmic bias. To discern algorithmic bias from dataset biases, it is essential to maintain a roughly equal distribution of other non-protected attributes, like background, lighting, and pose, across different social groups. Achieving this balance is impossible with observational data – that we cannot manipulate or intervene in – and all existing datasets fall short in meeting this requirement.

Synthetic datasets. Synthetic data generation has been a key focus in deep learning, addressing challenges in annotating data for tasks like image segmentation [40, 65], optical flow estimation [14, 56], OCR [30, 35, 43], and information retrieval [13, 22, 36]. Synthetic data not only reduces manual annotation efforts but also enriches training datasets, which is vital when real-world data acquisition is impractical or privacy-compromised.

Synthetic data has also been instrumental in creating controlled environments for model evaluation. For example, the CLEVR dataset [38], systematically generates images and questions that rigorously test various aspects of visual reasoning in VQA. Similarly, the bAbI QA tasks [79] are designed to test language understanding and reasoning via question answering. This demonstrates how synthetic data can provide highly structured and controlled settings, enabling the detailed assessment of models' capabilities.

In the specific domain of fairness evaluation using synthetic data, Balakrishnan *et al.* [3] and Liang *et al.* [50] used synthetically generated faces to benchmark face analysis and recognition

systems respectively, revealing lower accuracy for certain demographic subgroups. The use of synthetic data in these works addresses several challenges inherent in sampling real-world datasets. Obtaining a sufficiently large number of individuals for each protected demographic group is difficult when relying on real data. Moreover, ensuring an equal distribution of unprotected attributes (such as background, lighting, pose, etc.) across different groups is crucial to avoid misinterpreting dataset biases as algorithmic biases. This level of control is nearly impossible with observational data, which lacks the ability to be manipulated for such specific needs. By using a state-of-the-art Generative Adversarial Network (GAN) with explicit control over geometry and pose they were able to provide a more accurate and fair assessment of algorithmic performance across diverse demographic groups. In this paper, we take inspiration from prior works leveraging synthetic data for fairness evaluation, but focus on text-image retrieval systems.

3 Social bias assessment in text-image retrieval

Our bias assessment framework is built around a reference text-image dataset $\mathcal{D} = (q_1, \mathcal{I}_1), (q_2, \mathcal{I}_2), \ldots, (q_N, \mathcal{I}_N)$, where each q_i represents a textual query, and $\mathcal{I}_i = I_i^1, I_i^2, \ldots, I_i^M$ is a set of M images relevant to q_i. This dataset is structured with two specific requirements: (1) all textual queries (q_i) must use gender-neutral language, and (2) image sets ($\mathcal{I}_i$) must show a balanced representation of different demographic groups in terms of gender and ethnicity. More formally, given a set of protected attribute values $A = \{a_1, \ldots, a_l\}$ (for gender and ethnic demographic groups), $\mathcal{I}_i$ must have $\frac{M}{l}$ images having each possible attribute value in A. Details on how we build such a dataset are provided in the next section.

State-of-the-art text-image retrieval models learn embedding functions ϕ_q for the text queries and ϕ_I for the images. These functions project their respective inputs into a shared embedding space, aiming to position the representation of a text query $\phi_q(q)$ and an image $\phi_I(I)$ closely together if, and only if, the image I is relevant to the query q. The similarity between embeddings, quantified using a similarity metric $sim(\phi_q(q), \phi_I(I))$, guides the retrieval of relevant images in response to a given text query. Let $R(q, \mathcal{I})$ denote the retrieval ranking for a query q obtained as:

$$R(q, \mathcal{I}) = \text{argsort}_{I \in \mathcal{I}}[sim(\phi_q(q), \phi_I(I))] \tag{1}$$

where argsort is a sorting operation that orders the images in $\mathcal{I}$ in descending order by their similarity scores.

To assess the fairness of these rankings, we can leverage our dataset $\mathcal{D}$. Considering the balanced demographic representation in each subset $\mathcal{I}_i$, we apply the model to generate a retrieval ranking $R(q_i, \mathcal{I}_i)$ for each query q_i. An inclusive and fair model should produce a ranking that mirrors the demographic balance of $\mathcal{I}_i$. We evaluate this using a fairness metric based on Kullback-Leibler divergence, comparing the model's ranking against the expected uniform distribution of protected attributes in $\mathcal{I}_i$.

Consider two discrete distributions: $D_{R(q_i,\mathcal{I}_i)[:n]}$ and D_r. The former, $D_{R(q_i,\mathcal{I}_i)[:n]}$, assigns to each attribute value in A the proportion of images with that value within the top n images of the ranked list $R(q_i, \mathcal{I}_i)$. Conversely, D_r corresponds to the desired uniform distribution of these attribute values. By computing the KL-divergence between these distributions at each list position n, we measure the deviation from the desired attribute distribution.

We adopt a normalized discounted cumulative form of this metric [29, 64, 82]. This results in a non-negative metric where higher values indicate a greater disparity between $D_{R(q_i,\mathcal{I}_i)}$ and D_r. A zero value, the ideal case, reflects identical distributions at every rank position n. Formally, the normalized discounted cumulative KL-divergence (NDKL) for $R(q_i, \mathcal{I}_i)$ is defined as follows:

$$NDKL(R(q_i, \mathcal{I}_i)) = \frac{1}{Z} \sum_{n=1}^{|R(q_i,\mathcal{I}_i)|} \frac{dKL(D_{R(q_i,\mathcal{I}_i)[:n]} || D_r)}{\log_2(n+1)} \tag{2}$$

were $dKL(D1 \,||\, D2) = \sum_j D1(j) \log_e \left(\frac{D1(j)}{D2(j)}\right)$ is the KL-divergence between distributions $D1$ and $D2$. The normalization term Z is defined as $Z = \sum_{n=1}^{|R(q_i,\mathcal{I}_i)|} \frac{1}{\log_2(n+1)}$, ensuring the metric is scaled appropriately for comparison across different queries and rankings. Compared with other existing fairness metrics [64], $NDKL$ has the advantage of allowing multinomial protected attributes and soft group association, enabling a comprehensive assessment of fairness beyond binary attributes.

4 SISPI Dataset

Creating a real dataset that meets our social bias assessment framework's criteria is challenging, requiring meticulous manual work and making it difficult to find suitable images for certain queries. To overcome these challenges and ensure a diverse and robust dataset, we leverage state-of-the-art image generation models. These models can produce highly realistic images from textual descriptions, proving valuable for dataset creation across various tasks [10, 17, 51, 83]. For our dataset, named SISPI (Social Inclusive Synthetic Professionals Images), we used the Stable Diffusion XL model [62].

To construct the text prompts for image generation, we use a template prompt: "`An image of a <ETHNICITY> <GENDER> <PROFESSION>`". The placeholders `<ETHNICITY>` `<GENDER>` and `<PROFESSION>` are substituted respectively with words from ethnicity and gender protected attribute sets ({"asian", "black", "latin", "white"} and {"male", "female"}) and from a comprehensive list of 194 professions provided by Saunders *et al.* [66] (e.g. {"engineer", "assembler", "teacher", ...}), ensuring a broad and inclusive representation of various roles, occupations, and activities. This systematic approach enables us to generate a diverse and balanced set of images for each profession, aligning with the fair and balanced criteria of our framework.

In total the SISPI dataset comprises 49,664 images: for each of the 194 professions in the list collected by Saunders *et al.* [66], we create 8 prompt versions with all possible combinations of gender and ethnicity protected attributes' values, and generate 32 images with each prompt ($194 \times 8 \times 32 = 49,664$). Crucially, for each set of 8 images corresponding to a given profession, we use the same seed to initialize the image generation process. This approach allows us to precisely control the protected attributes – i.e., ethnicity and gender – while keeping the unprotected attributes, such as background settings, human poses, and object contexts, consistent across each set. This method ensures that variations in the images are attributable primarily to the protected attributes, rather than

Figure 2: Samples of the SISPI dataset for social bias assessment in text-image retrieval systems. For each text query we show 8 relevant images that originate from the same seed.

unrelated visual artifacts. Finally, the 194 queries (one per profession) are in the same form as the prompts used to generate the images, but protecting (hiding) the gender and ethnicity protected attributes' values. Figures 1 and 2 illustrate some examples of the SISPI dataset. The dataset and code of our framework is publicly available at https://sispi-benchmark.github.io/.

5 Experiments

In this section, we evaluate text-image retrieval models using our bias assessment framework with the SISPI dataset and the NDKL metric. The experiments are divided into two main parts. The first part assesses fairness across various CLIP-based models, which differ in size, training data, and methodology. The second part expands the analysis to include other leading retrieval architectures, providing a broader perspective on their fairness performance on the SISPI dataset. Following these evaluations, we present an in-depth discussion of the results, highlighting key insights, and analyzing specific queries qualitatively. For a detailed statistical significance analysis and additional qualitative examples, please refer to the supplementary material.

5.1 Comparing fairness of CLIP-based models

In this experiment, we apply the SISPI-NDKL metric to various CLIP models, including the original CLIP [63] and its extensions OpenCLIP [20], DFN [25], MetaCLIP [33], SigLIP [87], CLIPA [49], and EVA-CLIP [73]. These models vary significantly in architecture, training datasets, and methods, offering a comprehensive view of how these factors influence social bias in text-image retrieval. Our analysis seeks to reveal key differences in fairness performance across these models, highlighting the impact of their unique characteristics.

Table 1 provides performance metrics for various CLIP-based models of similar capacity, covering both standard retrieval metrics and fairness assessments on the SISPI dataset. We include Normalized Discounted Cumulative Gain (NDCG) and Mean Average Precision (mAP) to measure general retrieval performance, averaged over 194 queries using the full set of 49,664 images.

In addition to retrieval metrics, we report *NDKL* fairness results, focusing on gender, ethnicity, and their intersectional distribution. Unlike NDCG and mAP, *NDKL* is computed specifically for image subsets relevant to each of the 194 queries ($R(q_i, \mathcal{I}_i)$), allowing a detailed evaluation of how each model variant handles protected attribute distributions within its relevant results.

Overall, the NDKL fairness results suggest that all models are far from the worst-case scenario (extreme unfairness), though there is a clear margin for improvement. A notable gap of 0.1 in NDKL values exists between the fairest and most unfair models (color-coded blue and red), particularly in gender and intersectional gender-ethnicity distributions (0.08 vs. 0.19 and 0.29 vs. 0.39, respectively).

We appreciate that models sharing the same pre-training dataset tend to exhibit similar fairness outcomes, which is expected since biases in the dataset are propagated and often amplified during training, leading to biased outputs. Figure 4 shows the NDKL distribution by pre-training dataset, highlighting that models pretrained on WIT-400M [63] and MetaCLIP [33] outperform others in fairness. Unfortunately, the original CLIP paper [63] provides limited details on WIT-400M's collection, leading to attempts to recreate CLIP's data [34, 68, 69]. Metadata-Curated Language-Image Pre-training (MetaCLIP) [33] aims to reveal CLIP's data curation approach by leveraging a raw data pool and metadata (derived from CLIP's concepts) to create a balanced subset over the metadata distribution. In terms of its performance in the SISPI dataset this curation framework provides a clear advantage in comparison to other raw/uncurated datasets.

While averaging NDKL across queries provides a broad view of model performance, it can obscure important variations. A model may excel in fairness on some queries but show bias on others, with averaging masking both its strengths and weaknesses. This observation highlights the need for diverse evaluation strategies in fairness assessment. While average NDKL offers a useful summary, it should be complemented with more granular analyses. To this end, we conduct a query-by-query analysis to identify conditions where models perform more or less fairly.

Figure 3 presents a query-by-query analysis of models trained on three datasets: WIT-400M, MetaCLIP, and LAION400M. We selected

Table 1: Comparative performance for CLIP variants on the SISPI Dataset. We show the average NDCG and mAP, along with the average NDKL fairness metric for gender, ethnicity, and their joint distribution. NDKL values are color-coded to highlight the fairest (blue) and most unfair (red) models.

Model	Pre-training	NDCG ↑	mAP ↑	NDKL ↓ (gender)	NDKL ↓ (ethnic)	NDKL ↓
ResNet50	WIT-400M [63]	0.77	0.37	0.13	0.17	0.33
ResNet50x4		0.77	0.39	0.12	0.17	0.32
ResNet50x16		0.77	0.38	0.09	0.18	0.30
ResNet50x64		0.78	0.39	0.10	0.17	0.31
ResNet101		0.76	0.37	0.13	0.16	0.33
ViT-B-32		0.75	0.34	0.12	0.17	0.32
ViT-B-16		0.79	0.41	0.12	0.20	0.35
ViT-L-14		0.78	0.40	**0.08**	0.18	0.30
ViT-B-32	LAION-400M [69]	0.80	0.42	0.18	0.16	0.36
ViT-B-16		0.81	0.46	0.17	0.18	0.38
ViT-B-16-plus-240		0.82	0.48	0.16	0.16	0.35
ViT-L-14		0.82	0.48	0.16	0.16	0.34
ConvNeXt$_{base}$		0.80	0.44	0.17	0.19	0.39
EVA01-G-14		0.83	0.50	0.16	0.17	0.36
ViT-B-32	LAION-2B [68]	0.82	0.48	0.19	0.17	0.38
ViT-B-16		0.83	0.49	0.18	0.17	0.37
ViT-L-14		0.84	0.51	0.15	0.16	0.34
ViT-H-14		0.83	0.50	0.13	0.17	0.34
ViT-H-14-CLIPA-336		0.84	0.52	0.18	0.17	0.38
ViT-G-14		0.82	0.48	0.15	0.17	0.34
ViT-bigG-14		0.84	0.51	0.14	0.16	0.33
RoBERTa-ViT-B-32		0.82	0.47	0.19	0.15	0.37
ConvNeXt$_{base}$		0.83	0.49	0.16	0.18	0.37
ConvNeXt$_{large}$		0.84	0.51	0.18	0.18	0.38
ConvNeXt$_{xxlarge}$		0.85	0.53	0.17	0.17	0.37
EVA02-E-14		0.84	0.52	0.15	0.17	0.36
XLM-RoBERTa$_b$-ViT-B-32	LAION-5B [68]	0.81	0.46	0.17	0.18	0.38
XLM-RoBERTa$_l$-ViT-H-14		0.83	0.50	0.17	0.17	0.36
ViT-B-32	DataComp-1B [26]	0.80	0.45	0.14	**0.14**	0.31
ViT-B-16		0.82	0.48	0.16	0.16	0.34
ViT-L-14		0.82	0.48	0.18	0.16	0.36
ViT-L-14-CLIPA		0.83	0.51	0.19	0.16	0.37
ViT-H-14-CLIPA		0.84	0.51	0.17	0.16	0.36
ViT-B-32	MetaCLIP(400M) [33]	0.81	0.46	0.13	**0.14**	0.31
ViT-B-16		0.82	0.47	0.15	**0.14**	0.33
ViT-L-14		0.80	0.44	0.11	0.15	0.30
ViT-B-32	MetaCLIP(2.5B) [33]	0.82	0.47	0.15	0.15	0.33
ViT-B-16		0.81	0.47	0.15	0.15	0.32
ViT-L-14		0.83	0.49	0.12	0.15	0.30
ViT-H-14		0.83	0.49	0.11	0.15	**0.29**
ViT-B-16-SigLIP	WebLi [18]	0.80	0.44	0.18	0.15	0.36
ViT-L-16-SigLIP		0.82	0.47	0.16	0.17	0.35
ViT-SO400M-14-SigLIP		0.82	0.49	0.14	0.17	0.34
ViT-B-16	DFN2B [25]	0.82	0.48	0.18	0.15	0.35
ViT-L-14		0.83	0.50	0.15	0.16	0.34
ViT-H-14	DFN5B [25]	0.84	0.52	0.16	0.16	0.35

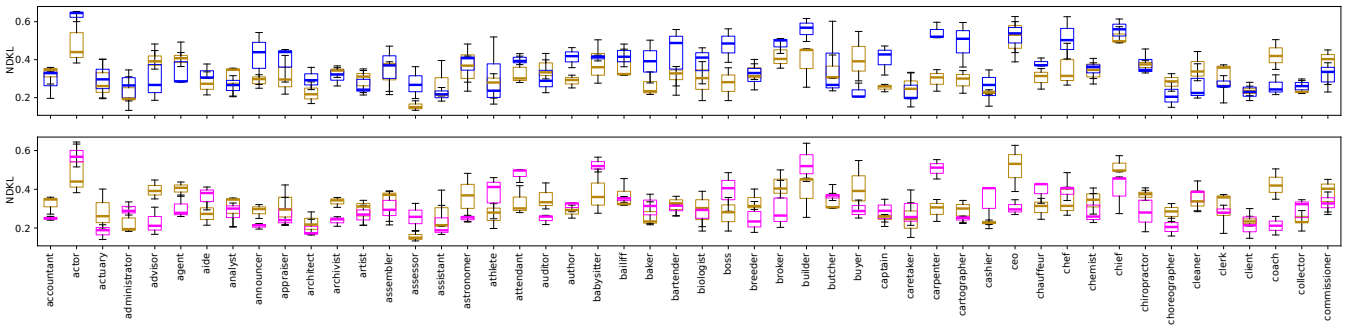

Figure 3: Query-by-Query Fairness Analysis for Models Trained on WIT-400M, MetaCLIP, and LAION400M Datasets. This figure presents boxplots of NDKL values across 50 selected queries.

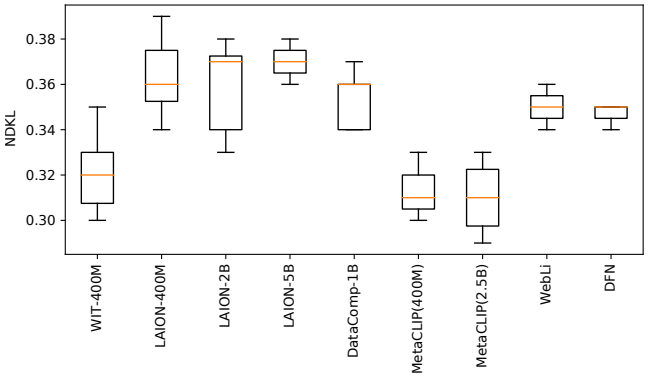

Figure 4: Distribution of NDKL fairness metric values of CLIP models by pre-training dataset. The concentration of similar fairness levels within models trained on identical datasets underscores the impact of pre-training data on bias in model outputs.

three common architectures – VIT-B-32, VIT-B-16, and VIT-L-14 – for consistent comparison. For clarity, results for 50 queries are shown here, while the rest can be found in the supplementary materials.

The analysis reveals significant fluctuations in NDKL values across queries, highlighting the variable nature of bias in text-image retrieval models. Models pretrained on WIT-400M generally show better fairness than those on LAION-400M, indicating potential inherent advantages in WIT-400M's dataset. However, LAION-trained models outperform WIT-trained ones in specific queries, such as "buyer," "choreographer," and "coach." The comparison between WIT and MetaCLIP-trained models shows no clear winner in fairness. WIT-trained models excel in some cases, while MetaCLIP-trained models lead in others. This parity highlights that both datasets have unique strengths and weaknesses, affecting fairness differently across queries.

5.2 Comparing fairness of state-of-the-art text-image retrieval models

In this subsection, we compare CLIP with other leading text-image retrieval models, including ALIGN [37], BLIP [45, 46], FLAVA [72], BridgeTower [81], and COCA [85]. These models differ in their architecture, training methods, and data.

Table 2 compares state-of-the-art models on standard retrieval metrics (NDCG, mAP) and fairness (NDKL) using the SISPI dataset. Notably, FLAVA models excel in NDKL fairness, while BridgeTower models match the best CLIP models. This suggests that FLAVA and BridgeTower models generally show less gender and ethnic bias. However, a critical consideration at this point is the retrieval performance of these models. FLAVA and BridgeTower have notably lower NDCG and mAP values. This poor performance may lead to models appearing fairer than they are, as random rankings can accidentally score well on fairness metrics like NDKL due to their statistical parity. Thus, fairness metrics should be interpreted alongside overall model performance, as random models may achieve high fairness scores by chance rather than actual bias reduction.

To illustrate this point more clearly, Figure 5 shows a scatter plot of NDCG versus NDKL values for various models evaluated in the SISPI dataset. This includes models from Tables 1 and 2, as well as additional CLIP models trained on smaller datasets and intermediate checkpoints from DataComp and CommonPool.

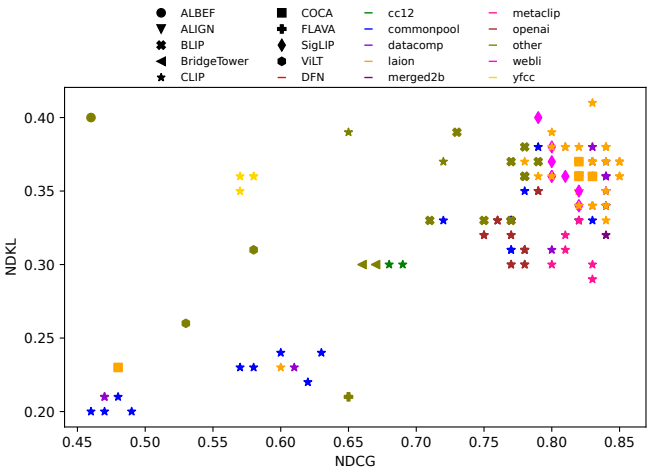

Figure 5: Contrasting retrieval utility (NDCG) and fairness (NDKL) in all evaluated models. Each point represents a model, with its position indicating its performance in terms of NDCG on the x-axis and NDKL on the y-axis.

Table 2: Comparative performance for state-of-the-art text-image retrieval models on the SISPI Dataset. We show the average NDCG and mAP, along with the average NDKL fairness metric for gender, ethnicity, and their joint distribution. CC: Conceptual Captions, SBU: SBU Captions, COCO: MSCOCO Captions, VG: Visual Genome. †FLAVA model was pretrained on public available 70M image and text pairs (including CC, VG, SBU, RedCaps, Wikipedia, and a custom filtered subset of YFCC100M).

Model	Pre-training	NDCG ↑	mAP ↑	NDKL ↓ (gender)	NDKL ↓ (ethnic)	NDKL ↓
CLIP ViT-L-14	WIT-400M	0.78	0.40	0.08	0.18	0.30
ALIGN	COYO-700M[15]	0.82	0.47	0.18	0.15	0.36
$BLIP_{Base}$	CC + SBU + COCO + VG + LAION-127M	0.77	0.38	0.18	0.17	0.37
$BLIP_{Large}$		0.79	0.42	0.20	0.15	0.37
$BLIP2_{Base}$		0.78	0.41	0.18	0.17	0.38
$BLIP2_{Large}$		0.78	0.40	0.15	0.17	0.36
FLAVA	†	0.65	0.19	**0.06**	**0.11**	**0.21**
$BridgeTower_{Base}$	CC + SBU + COCO + VG	0.66	0.22	0.13	0.13	0.30
$BridgeTower_{Large}$		0.67	0.25	0.14	0.12	0.30
COCA ViT-B-32	LAION-2B	0.82	0.47	0.18	0.16	0.37
COCA ViT-L-14		0.83	0.50	0.17	0.16	0.36

Figure 5 clearly illustrates two key findings of our study thus far: (1) the impact of training data on model fairness, and (2) the trade-off between retrieval effectiveness and fairness–highlighting the challenge of balancing high performance with fairness.

Regarding the impact of training data, we observe that models trained with the same dataset (points of the same color) tend to cluster together in the plot. This clustering indicates that the choice of training data significantly influences both retrieval performance and fairness metrics. Notice that for datasets such as "commonpool" (blue),"datacomp"(dark violet) and "laion" (orange), distinct clusters emerge that correspond to different dataset scales (small, medium, large, x-large). On the other hand, the "other" datasets (olive green) form a broad category where clustering is not expected due to their diverse origins.

Finally, to better understand the observed trade-off between retrieval metrics (e.g., NDCG and mAP) and fairness (NDKL), we consider an extreme hypothetical case: a retrieval model that assigns random vectors to each query and image. Such a model would exhibit very poor retrieval performance but achieve an almost perfect NDKL fairness score, as all images would have an equal chance of appearing in any ranking position regardless of protected attributes (i.e. gender and ethnicity). This example illustrates that weaker retrieval models might appear more "fair" simply due to their lack of differentiation in attribute values. Based on this insight, the optimal model would be positioned closest to the bottom-right corner of the plot, balancing high retrieval performance with fairness. In our case, the CLIP ViT-H-14 model trained on MetaCLIP(2.5B) achieves the best balance, with an NDKL score of 0.29 and an NDCG of 0.83.

6 Discussion

While our experimental evaluation provides a quantitative perspective on fairness in text-image retrieval models, it is crucial to delve deeper into the qualitative aspects and practical implications of the findings. This section addresses key questions that arise from our analysis, offering concrete examples and actionable insights to better understand how biases manifest in retrieval models.

Do text-image retrieval models reproduce non-inclusive stereotypes?

To address this question, we analyze specific queries with high NDKL scores. Figure 6 presents the top-8 retrieved images using the CLIP ViT-H-14 pretrained on MetaCLIP-2.5B (our best model) for the queries "babysitter" (NDKL = 0.57), "secretary" (NDKL = 0.54), and "hunter" (NDKL = 0.53). The demographic distribution in the top-ranked images is noticeably skewed toward stereotypical representations, reinforcing the association of certain roles with specific genders and ethnicities. For instance, all top-8 retrieved images for "babysitter" and "secretary" depict women, whereas all images for "hunter" feature men. Furthermore, a majority of "hunter" and "secretary" images portray non-white individuals, while most "babysitter" images predominantly exclude black individuals.

In Table 3, we present a list of professions with the highest and lowest NDKL scores across attributes for two different models. Additional qualitative examples are provided in the supplementary material. Figure 7 shows the distribution of NDKL values across gender, ethnicity, and intersectional fairness for all evaluated models with $mAP > 0.4$. Overall, we appreciate consistent trends where certain professions exhibit higher bias across various models and training datasets, reinforcing the presence of deeply ingrained societal stereotypes in text-image retrieval systems. Although some of the evaluated models achieve lower average NDKL scores compared to others, our analysis indicates that all models exhibit significant biases for certain queries.

Does the text-to-image generative model introduce bias in the SISPI dataset?

We generated images for different demographic groups using the same initial seeds to ensure that the joint distribution of unprotected attributes remained approximately equal across groups. This approach effectively mitigates demographic artifacts and biases potentially introduced by SDXL, ensuring that performance variations primarily reflect controlled attributes. Upon manual inspection of

Figure 6: Qualitative results for CLIP VIT-H-14 model pretrained on MetaCLIP-2,5B. We show the top-8 retrieved results for the text queries "A photo of a babysitter" (top row), "A photo of a secretary" (top row), "A photo of a hunter" (bottom row). The NDKL values for these particular queries are 0.57, 0.54, **and** 0.53 **respectively.**

Table 3: Professions with the highest and lowest NDKL scores across attributes for two different CLIP models.

	CLIP ViT-H-14 MetaCLIP-2.5B			CLIP ViT-H-14 LAION-2B		
	(gender)	(ethnic)	(all)	(gender)	(ethnic)	(all)
more bias	secretary (0.44)	host (0.46)	babysitter (0.57)	miner (0.45)	bailiff (0.44)	secretary (0.66)
	babysitter (0.42)	helper (0.44)	secretary (0.54)	nurse (0.45)	author (0.39)	builder (0.60)
	installer (0.40)	dancer (0.38)	hunter (0.53)	secretary (0.45)	dancer (0.38)	roofer (0.57)
	hunter (0.39)	pensioner (0.33)	mover (0.53)	builder (0.42)	warden (0.35)	miner (0.56)
	nurse (0.36)	planner (0.31)	host (0.52)	farmer (0.41)	curator (0.35)	worker (0.56)
less bias	analyst (0.02)	roofer (0.07)	lawyer (0.16)	salesperson (0.02)	customer (0.08)	employee (0.18)
	collector (0.02)	electrician (0.07)	veterinarian (0.16)	editor (0.02)	magistrate (0.08)	fundraiser (0.18)
	client (0.02)	lawyer (0.07)	paramedic (0.16)	representative (0.02)	legislator (0.08)	representative (0.18)
	teacher (0.02)	optician (0.06)	analyst (0.16)	professor (0.02)	chiropractor (0.07)	lawyer (0.17)
	bartender (0.02)	pathologist (0.06)	psychiatrist (0.15)	performer (0.02)	logistician (0.07)	paramedic (0.14

the generated data, we removed a few professions where the generator produced skewed stereotypical representations for certain demographic groups.

As part of the quality control process during the generation stage, we used the SDXL configuration for photo-realistic images (excluding illustrations, etc.). Additionally, we evaluated the quality of the synthetic data across different demographic groups using an aesthetic prediction method [74] trained on the Aesthetic Visual Analysis dataset [59]. The results showed no significant differences across groups, e.g. with scores of 6.61 for males' images and 6.47 for females' images on a scale of 1 to 10.

Do SISPI results generalize to real data?

While the SISPI dataset provides a controlled environment for evaluating fairness in text-image retrieval models, the extent to which these results generalize to real-world data requires careful consideration. SISPI is synthetically generated to ensure demographic balance and control over unprotected attributes, allowing for precise bias measurement. However, real-world data is inherently more complex, with uncontrolled factors such as cultural influences, regional variations, noise, and inherent biases present in web-scraped datasets.

We acknowledge the gap between synthetic and real-world data, a gap shared with other existing synthetic benchmarks [3, 38, 50, 79, 80] that have been instrumental in analyzing and improving the performance of machine learning models in different tasks. As observed in prior synthetic benchmarks, some degree of generalizability can be expected if the features leveraged by models to perform well on synthetic data are representative of those found in real-world data. In this work, we assume that current text-to-image models generate sufficiently realistic features to represent demographic groups as they appear in real data.

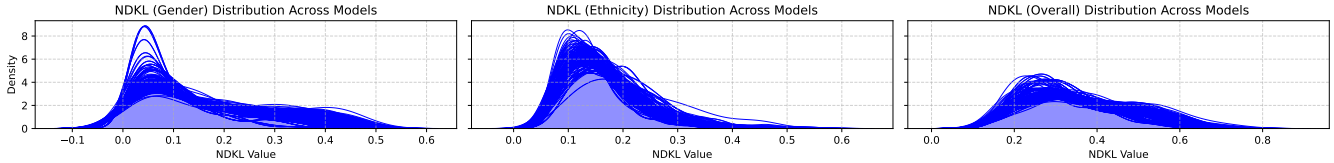

Figure 7: NDKL Value Distributions Across Evaluated Models. The figure presents the kernel density estimation (KDE) distributions of NDKL values for gender, ethnicity, and overall fairness across different models with $mAP > 0.4$. Each subplot represents the distribution of NDKL values across queries for the respective attribute.

To support this assumption, we evaluated a state-of-the-art fair face attribute classifier [39] on the SISPI dataset. Table 4 presents the obtained results. Since the classifier was trained on real-world data, its high accuracy across all demographic attributes suggests that SDXL generates sufficiently realistic features to represent demographic groups as they appear in real-world contexts.

Table 4: FairFace [39] Attribute Classifier Accuracy on the SISPI Dataset.

Metric	Accuracy
Overall Gender Accuracy	99.35%
Overall Race Accuracy	96.82%
Per-Class Gender Accuracy	
Male	99.36%
Female	99.34%
Per-Class Race Accuracy	
Asian	99.39%
White	94.82%
Black	98.94%
Latin	94.12%

Furthermore, our qualitative analysis indicates that biases observed in SISPI (e.g., overrepresentation of certain demographics in professional roles such as "babysitter", or "hunter") are recognizable reflections of societal stereotypes that also manifest in real-world datasets, albeit with additional noise and variability. These findings demonstrate that SISPI serves as a valuable benchmarking tool, but caution should be exercised when extrapolating results to real-world applications.

Additionally, our synthetic approach (akin to other synthetic benchmarks) allows to isolate model behavior regarding specific attributes without interference from other variables, thus offering controlled insights not achievable with real data, though real-world datasets indeed exhibit more complex biases.

In summary, we stress that fairness on SISPI is not an end goal in itself. SISPI provides a reliable framework for comparative fairness assessment across models, yet its findings should be complemented with evaluations on real-world datasets.

7 Ethical Considerations

In this study, we have adopted broad ethnic categories – "Asian," "White," "Black," and "Latin" – and gender categories of "Male" and "Female." While these are common in demographic research for their simplicity, they inherently oversimplify complex identities.

Ethnic Categorization: These categories encompass diverse cultures and histories, and terms like "Asian" oversimplify the rich diversity within each group. They also vary in perception and definition across regions.

Gender Categorization: The binary gender categories used here do not capture all gender identities. We acknowledge and respect non-binary and transgender identities.

Cultural Sensitivity and Inclusivity: We approach these classifications with sensitivity and acknowledge their limitations. Individuals' self-identification may be more nuanced, and we are open to feedback for improving our practices.

We made efforts to adhere to ethical practices in dataset generation by using consistent initial seeds for different demographic groups, ensuring a roughly equal distribution of unprotected attributes. This approach aims to minimize demographic artifacts and dataset biases. However, we acknowledge that the text-to-image generation model used may still introduce unintended biases that we have not been able to detect through manual inspection.

8 Conclusions

Our work introduces a novel framework for assessing gender and ethnicity bias in cross-modal text-image retrieval methods and underscores the value of synthetic datasets like SISPI for evaluating fairness. Our framework provides insights that are not attainable from previously existing frameworks that are based on real data.

In our experiments, we found that the choice of pre-training data significantly affects bias propagation, highlighting the need for careful dataset curation to ensure equitable model behavior. Our analysis also reveals trade-offs between retrieval metrics and fairness measures. Finally, we show that while average NDKL offers valuable insights, more detailed analyses – including NDKL distributions and query-by-query evaluations and qualitative results – provide a deeper understanding of model performance. Overall, our analysis indicates that all evaluated text-image retrieval models exhibit significant bias for certain queries, reproducing deeply ingrained non-inclusive stereotypes.

In summary, our research advocates for a nuanced, multi-faceted approach to fairness evaluation in cross-modal retrieval. By making our dataset and code publicly available, we aim to set the stage for future research into unbiased retrieval systems.

Acknowledgments

To Sonia Ruiz for valuable discussions and guidance on the political and ethical aspects of social bias during the early stages of this project. Lluis Gomez is funded by the Ramon y Cajal research fellowship RYC2020-030777-I / AEI / 10.13039/501100011033.

References

[1] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. 2022. Flamingo: a visual language model for few-shot learning. *Advances in Neural Information Processing Systems* 35 (2022), 23716–23736.

[2] Abolfazl Asudeh, HV Jagadish, Julia Stoyanovich, and Gautam Das. 2019. Designing fair ranking schemes. In *Proceedings of the 2019 international conference on management of data*. 1259–1276.

[3] Guha Balakrishnan, Yuanjun Xiong, Wei Xia, and Pietro Perona. 2021. Towards causal benchmarking of biasin face analysis algorithms. In *Deep Learning-Based Face Analytics*. Springer, 327–359.

[4] Hritik Bansal, Da Yin, Masoud Monajatipoor, and Kai-Wei Chang. 2022. How well can text-to-image generative models understand ethical natural language interventions? *arXiv preprint arXiv:2210.15230* (2022).

[5] Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big?. In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*. 610–623.

[6] Shruti Bhargava and David Forsyth. 2019. Exposing and correcting the gender bias in image captioning datasets and models. *arXiv preprint arXiv:1912.00578* (2019).

[7] Abeba Birhane, Sanghyun Han, Vishnu Boddeti, Sasha Luccioni, et al. 2023. Into the LAION's Den: Investigating Hate in Multimodal Datasets. In *Thirty-seventh Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

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

[9] Abeba Birhane, Vinay Uday Prabhu, and Emmanuel Kahembwe. 2021. Multimodal datasets: misogyny, pornography, and malignant stereotypes. *arXiv preprint arXiv:2110.01963* (2021).

[10] Nitzan Bitton-Guetta, Yonatan Bitton, Jack Hessel, Ludwig Schmidt, Yuval Elovici, Gabriel Stanovsky, and Roy Schwartz. 2023. Breaking Common Sense: WHOOPS! A Vision-and-Language Benchmark of Synthetic and Compositional Images. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 2616–2627.

[11] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems* 29 (2016).

[12] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. 2021. On the opportunities and risks of foundation models. *arXiv preprint arXiv:2108.07258* (2021).

[13] Luiz Bonifacio, Hugo Abonizio, Marzieh Fadaee, and Rodrigo Nogueira. 2022. Inpars: Unsupervised dataset generation for information retrieval. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2387–2392.

[14] Daniel J Butler, Jonas Wulff, Garrett B Stanley, and Michael J Black. 2012. A naturalistic open source movie for optical flow evaluation. In *Computer Vision–ECCV 2012: 12th European Conference on Computer Vision, Florence, Italy, October 7-13, 2012, Proceedings, Part VI 12*. Springer, 611–625.

[15] Minwoo Byeon, Beomhee Park, Haecheon Kim, Sungjun Lee, Woonhyuk Baek, and Saehoon Kim. 2022. COYO-700M: Image-Text Pair Dataset. https://github.com/kakaobrain/coyo-dataset.

[16] Jaime Carbonell and Jade Goldstein. 1998. The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st annual international ACM SIGIR conference on Research and development in information retrieval*. 335–336.

[17] Paola Cascante-Bonilla, Khaled Shehada, James Seale Smith, Sivan Doveh, Donghyun Kim, Rameswar Panda, Gul Varol, Aude Oliva, Vicente Ordonez, Rogerio Feris, and Leonid Karlinsky. 2023. Going Beyond Nouns With Vision & Language Models Using Synthetic Data. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 20155–20165.

[18] Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, et al. 2022. PaLI: A Jointly-Scaled Multilingual Language-Image Model. In *The Eleventh International Conference on Learning Representations*.

[19] Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Uniter: Universal image-text representation learning. In *European conference on computer vision*. Springer, 104–120.

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

[21] Won Ik Cho, Ji Won Kim, Seok Min Kim, and Nam Soo Kim. 2019. On measuring gender bias in translation of gender-neutral pronouns. *arXiv preprint arXiv:1905.11684* (2019).

[22] Zhuyun Dai, Vincent Y Zhao, Ji Ma, Yi Luan, Jianmo Ni, Jing Lu, Anton Bakalov, Kelvin Guu, Keith Hall, and Ming-Wei Chang. 2022. Promptagator: Few-shot Dense Retrieval From 8 Examples. In *The Eleventh International Conference on Learning Representations*.

[23] Alessandro Fabris, Alberto Purpura, Gianmaria Silvello, and Gian Antonio Susto. 2020. Gender stereotype reinforcement: Measuring the gender bias conveyed by ranking algorithms. *Information Processing & Management* 57, 6 (2020), 102377.

[24] Fartash Faghri, David J Fleet, Jamie Ryan Kiros, and Sanja Fidler. 2017. Vse++: Improving visual-semantic embeddings with hard negatives. *arXiv preprint arXiv:1707.05612* (2017).

[25] Alex Fang, Albin Madappally Jose, Amit Jain, Ludwig Schmidt, Alexander T Toshev, and Vaishaal Shankar. 2023. Data Filtering Networks. In *NeurIPS 2023 Workshop on Distribution Shifts: New Frontiers with Foundation Models*.

[26] 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. 2023. DataComp: In search of the next generation of multimodal datasets. arXiv:2304.14108 [cs.CV]

[27] Noa Garcia, Yusuke Hirota, Yankun Wu, and Yuta Nakashima. 2023. Uncurated image-text datasets: Shedding light on demographic bias. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 6957–6966.

[28] Timnit Gebru. 2020. Race and gender. *The Oxford handbook of ethics of AI* (2020), 251–269.

[29] Sahin Cem Geyik, Stuart Ambler, and Krishnaram Kenthapadi. 2019. 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*. 2221–2231.

[30] Ankush Gupta, Andrea Vedaldi, and Andrew Zisserman. 2016. Synthetic data for text localisation in natural images. 2315–2324.

[31] Laura Gustafson, Chloe Rolland, Nikhila Ravi, Quentin Duval, Aaron Adcock, Cheng-Yang Fu, Melissa Hall, and Candace Ross. 2023. FACET: Fairness in computer vision evaluation benchmark. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 20370–20382.

[32] Lisa Anne Hendricks, Kaylee Burns, Kate Saenko, Trevor Darrell, and Anna Rohrbach. 2018. Women also snowboard: Overcoming bias in captioning models. In *Proceedings of the European conference on computer vision (ECCV)*. 771–787.

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

[34] 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. 2021. *OpenCLIP*. doi:10.5281/zenodo.5143773 If you use this software, please cite it as below..

[35] M Jaderberg, K Simonyan, A Vedaldi, and A Zisserman. 2014. Synthetic data and artificial neural networks for natural scene text recognition. In *NIPS Deep Learning Workshop*. Neural Information Processing Systems.

[36] Vitor Jeronymo, Luiz Bonifacio, Hugo Abonizio, Marzieh Fadaee, Roberto Lotufo, Jakub Zavrel, and Rodrigo Nogueira. 2023. InPars-v2: Large Language Models as Efficient Dataset Generators for Information Retrieval. *arXiv preprint arXiv:2301.01820* (2023).

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

[38] Justin Johnson, Bharath Hariharan, Laurens Van Der Maaten, Li Fei-Fei, C Lawrence Zitnick, and Ross Girshick. 2017. Clevr: A diagnostic dataset for compositional language and elementary visual reasoning. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 2901–2910.

[39] Kimmo Karkkainen and Jungseock Joo. 2021. Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*. 1548–1558.

[40] Samin Khan, Buu Phan, Rick Salay, and Krzysztof Czarnecki. 2019. ProcSy: Procedural Synthetic Dataset Generation Towards Influence Factor Studies Of Semantic Segmentation Networks.. In *CVPR workshops*, Vol. 3. 4.

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

[42] Klara Krieg, Emilia Parada-Cabaleiro, Gertraud Medicus, Oleg Lesota, Markus Schedl, and Navid Rekabsaz. 2022. Grep-BiasIR: a dataset for investigating gender representation-bias in information retrieval results. In *Proceeding of the 2023 ACM*

SIGIR Conference On Human Information Interaction And Retrieval (CHIIR).
[43] Praveen Krishnan and CV Jawahar. 2019. HWNet v2: an efficient word image representation for handwritten documents. *International Journal on Document Analysis and Recognition (IJDAR)* 22, 4 (2019), 387–405.
[44] Kuang-Huei Lee, Xi Chen, Gang Hua, Houdong Hu, and Xiaodong He. 2018. Stacked cross attention for image-text matching.
[45] Junnan Li, Dongxu Li, Silvio Savarese, and Steven Hoi. 2023. Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. In *International Conference on Machine Learning*. PMLR.
[46] Junnan Li, Dongxu Li, Caiming Xiong, and Steven Hoi. 2022. Blip: Bootstrapping language-image pre-training for unified vision-language understanding and generation. In *International Conference on Machine Learning*. PMLR, 12888–12900.
[47] Junnan Li, Ramprasaath Selvaraju, Akhilesh Gotmare, Shafiq Joty, Caiming Xiong, and Steven Chu Hong Hoi. 2021. Align before fuse: Vision and language representation learning with momentum distillation. *Advances in neural information processing systems* 34 (2021), 9694–9705.
[48] Kunpeng Li, Yulun Zhang, Kai Li, Yuanyuan Li, and Yun Fu. 2019. Visual semantic reasoning for image-text matching. In *Proceedings of the IEEE International Conference on Computer Vision*. 4654–4662.
[49] Xianhang Li, Zeyu Wang, and Cihang Xie. 2023. An Inverse Scaling Law for CLIP Training. In *NeurIPS*.
[50] Hao Liang, Pietro Perona, and Guha Balakrishnan. 2023. Benchmarking Algorithmic Bias in Face Recognition: An Experimental Approach Using Synthetic Faces and Human Evaluation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 4977–4987.
[51] Hao Liang, Pietro Perona, and Guha Balakrishnan. 2023. Benchmarking Algorithmic Bias in Face Recognition: An Experimental Approach Using Synthetic Faces and Human Evaluation. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 4977–4987.
[52] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*. Springer, 740–755.
[53] Chunxiao Liu, Zhendong Mao, Tianzhu Zhang, Hongtao Xie, Bin Wang, and Yongdong Zhang. 2020. Graph structured network for image-text matching. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10921–10930.
[54] Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. *arXiv preprint arXiv:1908.02265* (2019).
[55] Jiasen Lu, Vedanuj Goswami, Marcus Rohrbach, Devi Parikh, and Stefan Lee. 2020. 12-in-1: Multi-task vision and language representation learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 10437–10446.
[56] Nikolaus Mayer, Eddy Ilg, Philip Hausser, Philipp Fischer, Daniel Cremers, Alexey Dosovitskiy, and Thomas Brox. 2016. A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 4040–4048.
[57] Nicole Meister, Dora Zhao, Angelina Wang, Vikram V Ramaswamy, Ruth Fong, and Olga Russakovsky. 2023. Gender artifacts in visual datasets. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 4837–4848.
[58] Margaret Mitchell, Dylan Baker, Nyalleng Moorosi, Emily Denton, Ben Hutchinson, Alex Hanna, Timnit Gebru, and Jamie Morgenstern. 2020. Diversity and inclusion metrics in subset selection. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. 117–123.
[59] Naila Murray, Luca Marchesotti, and Florent Perronnin. 2012. AVA: A large-scale database for aesthetic visual analysis. In *2012 IEEE conference on computer vision and pattern recognition*. IEEE, 2408–2415.
[60] Jahna Otterbacher, Alessandro Checco, Gianluca Demartini, and Paul Clough. 2018. Investigating user perception of gender bias in image search: the role of sexism. In *The 41st International ACM SIGIR conference on research & development in information retrieval*. 933–936.
[61] Christine Pinney, Amifa Raj, Alex Hanna, and Michael D Ekstrand. 2023. Much Ado About Gender: Current Practices and Future Recommendations for Appropriate Gender-Aware Information Access. In *Proceedings of the 2023 Conference on Human Information Interaction and Retrieval*. 269–279.
[62] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. 2023. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952* (2023).
[63] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PMLR, 8748–8763.
[64] Amifa Raj and Michael D Ekstrand. 2022. Measuring fairness in ranked results: An analytical and empirical comparison. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 726–736.
[65] German Ros, Laura Sellart, Joanna Materzynska, David Vazquez, and Antonio M Lopez. 2016. The synthia dataset: A large collection of synthetic images for semantic segmentation of urban scenes. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 3234–3243.
[66] Danielle Saunders and Bill Byrne. 2020. Reducing Gender Bias in Neural Machine Translation as a Domain Adaptation Problem. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 7724–7736.
[67] Beatrice Savoldi, Marco Gaido, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2021. Gender Bias in Machine Translation. *Transactions of the Association for Computational Linguistics* 9 (08 2021), 845–874.
[68] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. 2022. LAION-5B: An open large-scale dataset for training next generation image-text models. arXiv:2210.08402 [cs.CV]
[69] Christoph Schuhmann, Richard Vencu, Romain Beaumont, Robert Kaczmarczyk, Clayton Mullis, Aarush Katta, Theo Coombes, Jenia Jitsev, and Aran Komatsuzaki. 2021. LAION-400M: Open Dataset of CLIP-Filtered 400 Million Image-Text Pairs. arXiv:2111.02114 [cs.CV]
[70] Candice Schumann, Susanna Ricco, Utsav Prabhu, Vittorio Ferrari, and Caroline Pantofaru. 2021. A step toward more inclusive people annotations for fairness. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. 916–925.
[71] Piyush Sharma, Nan Ding, Sebastian Goodman, and Radu Soricut. 2018. Conceptual Captions: A Cleaned, Hypernymed, Image Alt-text Dataset For Automatic Image Captioning. In *Annual Meeting of the Association for Computational Linguistics*. 2556–2565.
[72] Amanpreet Singh, Ronghang Hu, Vedanuj Goswami, Guillaume Couairon, Wojciech Galuba, Marcus Rohrbach, and Douwe Kiela. 2022. Flava: A foundational language and vision alignment model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 15638–15650.
[73] Quan Sun, Yuxin Fang, Ledell Wu, Xinlong Wang, and Yue Cao. 2023. Eva-clip: Improved training techniques for clip at scale. *arXiv preprint arXiv:2303.15389* (2023).
[74] Hossein Talebi and Peyman Milanfar. 2018. NIMA: Neural image assessment. *IEEE transactions on image processing* 27, 8 (2018), 3998–4011.
[75] Hao Tan and Mohit Bansal. 2019. LXMERT: Learning Cross-Modality Encoder Representations from Transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 5100–5111.
[76] Ruixiang Tang, Mengnan Du, Yuening Li, Zirui Liu, Na Zou, and Xia Hu. 2021. Mitigating gender bias in captioning systems. In *Proceedings of the Web Conference 2021*. 633–645.
[77] Zihao Wang, Xihui Liu, Hongsheng Li, Lu Sheng, Junjie Yan, Xiaogang Wang, and Jing Shao. 2019. CAMP: Cross-Modal Adaptive Message Passing for Text-Image Retrieval.
[78] Laura Weidinger, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, et al. 2021. Ethical and social risks of harm from language models. *arXiv preprint arXiv:2112.04359* (2021).
[79] Jason Weston, Antoine Bordes, Sumit Chopra, Alexander M Rush, Bart Van Merriënboer, Armand Joulin, and Tomas Mikolov. 2016. Towards AI-complete question answering: A set of prerequisite toy tasks. In *4th International Conference on Learning Representations, ICLR 2016*.
[80] Depeng Xu, Shuhan Yuan, Lu Zhang, and Xintao Wu. 2018. Fairgan: Fairness-aware generative adversarial networks. In *2018 IEEE International Conference on Big Data (Big Data)*. IEEE, 570–575.
[81] Xiao Xu, Chenfei Wu, Shachar Rosenman, Vasudev Lal, Wanxiang Che, and Nan Duan. 2023. Bridgetower: Building bridges between encoders in vision-language representation learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 37. 10637–10647.
[82] Ke Yang and Julia Stoyanovich. 2017. Measuring fairness in ranked outputs. In *Proceedings of the 29th international conference on scientific and statistical database management*. 1–6.
[83] Zhitao Yang, Zhongang Cai, Haiyi Mei, Shuai Liu, Zhaoxi Chen, Weiye Xiao, Yukun Wei, Zhongfei Qing, Chen Wei, Bo Dai, Wayne Wu, Chen Qian, Dahua Lin, Ziwei Liu, and Lei Yang. 2023. SynBody: Synthetic Dataset with Layered Human Models for 3D Human Perception and Modeling. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 20282–20292.
[84] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics* 2 (2014), 67–78.
[85] Jiahui Yu, Zirui Wang, Vijay Vasudevan, Legg Yeung, Mojtaba Seyedhosseini, and Yonghui Wu. 2022. CoCa: Contrastive Captioners are Image-Text Foundation Models. *Transactions on Machine Learning Research* Aug 2022 (2022). https://arxiv.org/abs/2205.01917

[86] Meike Zehlike, Francesco Bonchi, Carlos Castillo, Sara Hajian, Mohamed Megahed, and Ricardo Baeza-Yates. 2017. Fa* ir: A fair top-k ranking algorithm. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*. 1569–1578.

[87] Xiaohua Zhai, Basil Mustafa, Alexander Kolesnikov, and Lucas Beyer. 2023. Sigmoid Loss for Language Image Pre-Training. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*. 11975–11986.

[88] Pengchuan Zhang, Xiujun Li, Xiaowei Hu, Jianwei Yang, Lei Zhang, Lijuan Wang, Yejin Choi, and Jianfeng Gao. 2021. Vinvl: Revisiting visual representations in vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 5579–5588.

[89] Dora Zhao, Angelina Wang, and Olga Russakovsky. 2021. Understanding and evaluating racial biases in image captioning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 14830–14840.

[90] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. *arXiv preprint arXiv:1707.09457* (2017).