Balancing the Picture: Debiasing Vision-Language Datasets with Synthetic Contrast Sets

Brandon Smith[*] **Miguel Farinha**[*] **Siobhan Mackenzie Hall**

Hannah Rose Kirk **Aleksandar Shtedritski** **Max Bain**

Oxford Artificial Intelligence Society, University of Oxford
https://github.com/oxai/debias-gensynth

Abstract

Vision-language models are growing in popularity and public visibility to generate, edit, and caption images at scale; but their outputs can perpetuate and amplify societal biases learned during pre-training on uncurated image-text pairs from the internet. Although debiasing methods have been proposed, we argue that these measurements of *model bias* lack validity due to *dataset bias*. We demonstrate there are spurious correlations in COCO Captions, the most commonly used dataset for evaluating bias, between background context and the gender of people in-situ. This is problematic because commonly-used bias metrics (such as Bias@K) rely on per-gender base rates. To address this issue, we propose a novel dataset debiasing pipeline to augment the COCO dataset with synthetic, gender-balanced contrast sets, where only the gender of the subject is edited and the background is fixed. As existing image editing methods have limitations and sometimes produce low-quality images; we introduce a method to automatically filter the generated images based on their similarity to real images. Using our balanced synthetic contrast sets, we benchmark bias in multiple CLIP-based models, demonstrating how metrics are skewed by imbalance in the original COCO images. Our results indicate that the proposed approach improves the validity of the evaluation, ultimately contributing to more realistic understanding of bias in CLIP.

1 Introduction

Vision-Language Models (VLMs) are rapidly advancing in capability and have witnessed a dramatic growth in public visibility with millions of daily active user of models such as DALL-E [1], Mid-Journey [2, 3], Stable Diffusion [4]. Underpinning these powerful generative models are image-text encoders like CLIP [5], which are themselves used for many discriminative tasks, such as video action recognition, open set detection and segmentation, and captioning. These encoders are pre-trained on large-scale internet scraped datasets. The uncurated nature of such datasets can translate to generated images that risk inflicting a range of downstream harms on their end users and society at large – from bias and negative stereotypes, to nudity and sexual content, or violent or graphic imagery [6, 7].

In light of these issues, coupled with growing use of generative AI, it is vital to reliably benchmark the bias in VLMs, particularly in the image-text encoders. A small emerging body of work attempts to measure bias in VLMs [8–10], or to debias their feature representations [9, 10]. Yet the legitimacy of this work critically depends on both a suitable evaluation metric and an evaluation dataset to

[*]Joint first authorship.

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

accurately depict the bias in pre-trained model weights and reliably signal whether debiasing attempts have been successful. The predominant focus on model-centric debiasing methods has overshadowed two main challenges associated with datasets and metrics: (i) the common use of cropped face datasets, such as FairFace [11], fall short because excluding contextual background presents an inaccurate and unreliable assessment of bias in natural images; and (ii) even if natural, open-domain images containing contextual clues are used, they are unbalanced by identity attribute representation within contexts. This is problematic because commonly-used bias metrics, such as Bias@K, are affected by the naturally-occurring distribution of images. Thus, while using contextual images is desirable, it comes at the cost of spurious correlations, affecting the reliability of bias metrics.

In this paper, we argue that these confounding factors arising from the interaction of metric choice and biased datasets paint an unreliable picture when measuring model bias in VLMs. To counter these issues, we propose a synthetic pipeline for debiasing a dataset into contrast sets balanced by identity attributes across background contexts. Our pipeline draws on the success of contrast sets in NLPs [12] and leverages recent advances in controllable image editing and generation [13]. We illustrate our approach with a focus on gender bias and define a contrast set as containing pairs of images from COCO [14] where each image ID has two synthetically-edited versions (one man, one woman) where the background is fixed and only the person bounding box is edited. We make three key contributions: (1) We demonstrate spurious correlations in the COCO dataset between gender and context, and show their problematic effects when used to measure model bias (Sec. 3); (2) We present the GENSYNTH dataset, built from a generative pipeline for synthetic image editing, and a filtering pipeline using KNN with real and synthetic images to control for the quality of the generated images (Sec. 4); (3) We benchmark CLIP models [5, 15] on our GENSYNTH dataset, which has no spurious correlation, and cast doubts on the effectiveness of debiasing methods (Sec. 5).

Our findings demonstrate that debiasing a dataset with synthetic contrast sets can avoid spurious correlations and more reliably measure model bias. While synthetically-edited data has promise in (i) preserving privacy of subjects included in vision datasets, and (ii) adding controllability to the dataset features, it also risks introducing a real-synthetic distribution shift and stacking biases of various generative models may essentialise representations of gender (see appendix A). Despite these early-stage limitations, this work starts a conversation about the importance of the interaction between dataset features with bias metrics, ultimately contributing to future work that paints a more accurate and balanced picture of identity-based bias in VLMs.

2 Related works

Defining Fairness and Bias. Fairness is a complex, context-dependent concept [16, 17]. Here, we adopt a narrow definition where no group is advantaged or disadvantaged based on the protected attribute of gender in retrieval settings [18, 19]. The metrics employed in this paper, *Bias@K* [20] and *Skew@K*, [21] are used to assess disparity in distribution between search results and desired outcomes. In this work, we assume activities such as *dancing, skateboarding, laughing* would not have a strong gendered prior and thus the desired distribution is one where all protected attributes have equal chance of being returned in a query that does not explicitly mention gender.[2]

Measuring Model Bias. Measuring bias in VLMs is a growing area of research. Early work measures classification rates of faces into harmful categories [8]. Several works measure outcome bias for text-to-face retrieval [9, 10, 22], though it is unclear how measurements made on cropped face datasets generalise to real-world settings. For gender fairness in open-domain images, COCO Captions [14] is a standard benchmark for cross-modal retrieval [20, 23] and image captioning [18, 24].

Dataset Bias. Datasets, including those used for bias evaluation, have their own biases from curation and annotation artefacts. Image datasets have been found to include imbalanced demographic representation [24–29], stereotypical portrayals [30–32], or graphic, sexually-explicit and other harmful content [6]. Similar to [29, 33], we find spurious correlations in COCO and show this makes it unsuitable for bias measurement. Methods to reduce dataset biases range from automatic [34] to manual filtering [35] of harmful images, such as those containing nudity [34], toxicity, or personal and identifiable information [36]. Yet, these filters cannot identify subtle spurious correlations present in open-domain images – making it difficult to curate an unbiased natural image dataset [33].

[2] In certain contexts, e.g. pregnant women, we may not want an equal distribution of masculine and feminine images to be returned, though we must not conflate biological gender and gender identity (see appendix A).

Mitigating Dataset Bias with Synthetic Data. Deep networks need large amounts of labeled data, prompting the creation of synthetic datasets for various computer vision tasks [37–40]. More recently, progress in generative models [1, 4, 41] has enabled methods to synthetically generate training data [13, 42–44]. Similarly, text-guided editing methods [13, 45, 46] offer scalable and controllable image editing, potentially enhancing dataset fairness. While synthetic datasets have been proposed to mitigate bias [19, 47], generated data may not represent underlying distributions of marginalised groups [48–51]. To combat these risks, fairness in generative models is an area gaining popularity [19, 52]. Similar to our work, GAN-based frameworks [53, 54] edit an *existing* face dataset to equalise attributes and enforce fairness. Our work extends this approach to open-domain images, introducing an automatic filtering technique for improving the quality of edits. Our work is also inspired by the use of contrast sets in NLP [12], which have been used to alter data by perturbing demographics (race, age, gender) to improve fairness [55]. We synthetically-generate contrast sets by augmenting both the textual and visual input to CLIP, for a more accurate evaluation of VLM bias.

3 Measuring Gender Bias on Natural Images

While prior works make in-depth comparisons between models, and even metrics [9], there is a dearth of research on whether natural image datasets, with their own biased and spurious correlations, are suitable benchmarks to measure bias in VLMs. In this section, we discuss the effect of dataset bias from data imbalance and spurious correlations on reliably measuring model bias.

3.1 Preliminaries

Bias measures. We look at two commonly used metrics – *Bias@K* [20] and *MaxSkew@K* [9, 21]. Bias@K measures the overrepresentation of men in the top K retrieval results. Skew@K measures the difference between the desired proportion of image attributes and the observed one, and MaxSkew@K is the maximum Skew among all attributes, or the "largest unfair advantage" [21] belonging to images of any perceived gender presentation. Bias@K is inherently a signed bias measure, which can be positive or negative for different samples, whereas MaxSkew@K is always positive. For further definitions and discussions of these, please refer to the Appendix. Our definition of bias measures a form of representational fairness, i.e., with a gender-balanced set of images and a gender-neutral caption, whether each perceived gender group has equal chances of being retrieved.

Data. To measure bias we use COCO – a dataset of 118k images with detection, segmentation and caption annotations [14, 56]. Each image has five captions written by different annotators. COCO is commonly used to measure gender bias in VLMs in tandem with the Bias@K metric [10, 20, 23].

3.2 Gendered Captions and Images in COCO

The bias metrics defined in Sec. 3.1 require gender attribute labels for each image and gender-neutral text queries, but these are not naturally present in captioned image data such as COCO. We describe the steps to automatically label gender for images and to neutralise gender information in captions.

Extracting Image Gender Labels from Captions. We assign a gender label to each COCO image, following prior work [20]. For each image, we concatenate all five captions into a single paragraph. If the paragraph contains only feminine words and no masculine words, the image is assigned a female label, and vice versa. If the paragraph contains words from both or neither genders, it is labeled as undefined. The full list of gendered words is detailed in the Appendix. Using this procedure, we implement the function g in Sec. 3.1. The COCO 2017 validation set contains 5,000 images, of which 1,275 (25.5%), are male, 539 (10.8%) female, and 3,186 (63.7%) undefined. This procedure gives high precision in the gender-pseudo label, as any ambiguous samples are rejected. However, images may be incorrectly labeled as undefined (lower recall) due to, for example, misspelling of the gendered words in the human-annotated captions or missing rare gendered terms in our keyword list.

Constructing Gender-Neutral Captions. We construct gender-neutral captions by replacing gendered words with neutral ones, e.g. "man" or "woman" become "person", and the sentence "A *man* sleeping with *his* cat next to *him*" becomes "A *person* sleeping with *their* car next to *them*". The full mapping of gender-neutral words and more examples are in the Appendix.

Table 1: Comparison of model gender bias for CLIP [5], a theoretically gender-agnostic model (TF-IDF on non-gendered words) and a random model, on the COCO validation set under unbalanced and balanced (with standard deviation computed over 5 runs) settings.

Model	COCO Val				COCO Val (Balanced)			
	Bias@K		MaxSkew@K		Bias@K		MaxSkew@K	
	K=5	K=10	K=25	K=100	K=5	K=10	K=25	K=100
Random Model	0.37	0.40	0.15	0.06	$0.00_{\pm 0.07}$	$0.00_{\pm 0.07}$	$0.14_{\pm 0.00}$	$0.07_{\pm 0.00}$
TF-IDF$_{\text{gender}-\text{agnostic}}$	0.22	0.24	0.29	0.22	$-0.06_{\pm 0.00}$	$-0.08_{\pm 0.00}$	$0.25_{\pm 0.00}$	$0.18_{\pm 0.00}$
CLIP	0.20	0.23	0.28	0.23	$-0.03_{\pm 0.01}$	$-0.06_{\pm 0.01}$	$0.24_{\pm 0.00}$	$0.19_{\pm 0.01}$

3.3 The Effect of Dataset Bias on Model Bias Measurement

There are two angles to dataset bias: (1) over representation – there are more photos of men than women in COCO on average, and (2) spurious correlations, i.e., some environments/backgrounds are more prevalent for specific gender groups. Accordingly, the dataset used for bias evaluation significantly affects the model bias measurement. This is exemplified by a theoretically gender-agnostic model, which we instantiate as a TF-IDF (Term Frequency - Inverse Document Frequency) ranking model for caption-to-caption retrieval on gender-neutralised captions. Despite being based on a simple numerical statistic of word occurrences, devoid of any inherent gender bias, this model still exhibits non-zero bias when evaluated on COCO captions. Our findings, reported in Tab. 1, include Bias@K and MaxSkew@K measurements on COCO Val, compared against a random model and CLIP. For Balanced COCO Val, all models register an approximate Bias@K of zero, a consequence of the metric's signed nature that tends to average towards zero over many directions of spurious correlations on biased but balanced data. Yet, for unbalanced data, Bias@K shifts towards the over-represented attribute in the dataset, making it an unsuitable metric for model bias measurement, as it reflects dataset bias instead. MaxSkew@K, despite being an absolute measure, is not exempt from these issues. It still records large values for the theoretically gender-agnostic model and the random model, suggesting that the established framework may be inadequate for bias measurement on natural image datasets that inherently possess their own biases. The experiments in Tab. 1 show the inadequacy of Bias@K and MaxSkew@K for measuring bias on natural image datasets that are imbalanced. Even when correcting for the imbalance issue, spurious correlations remain in the data, as we show in the Appendix. Therefore, we argue that a balanced unbiased dataset is required to robustly measure bias and compare different pre-trained models and debiasing strategies. This motivated us to propose GENSYNTH in Sec. 4, which attempts to fix both issues.

4 GENSYNTH: A Synthetic Gender-Balanced Dataset using Contrast Sets

Given the limitations of measuring Bias@K and MaxSkew@K on natural images and the spurious correlations in existing datasets, we propose a framework for editing natural images into *synthetic*

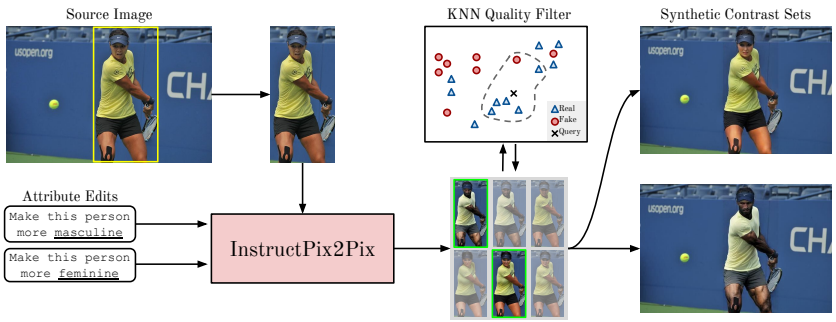


Figure 1: An overview of our pipeline for *dataset debiasing* across a target attribute, in this case gender, ensuring equal demographic representation. A source image containing a person is given as input to InstructPix2Pix along with instructions to synthesise each attribute label. The resulting edits are filtered for quality via K-Nearest Neighbour (KNN) thresholding to ensure realistic-looking edits for each attribute label (male and female).

contrast sets that remove spurious background correlations along the attribute of interest and apply the pipeline on COCO to obtain the GENSYNTH dataset (see Fig. 1). We first synthetically edit the person in images to cover both gender labels with fixed background context (Sec. 4.1), followed by automatic filtering that ensures the quality and correctness of the edited persons (Sec. 4.2). Finally, we verify the quality of the edited images and the filtering method (appendix F). While we implement this for the gender attribute, in practice, our pipeline could be used to generate synthetic contrast sets for other identity attributes, requiring only the availability of person bounding boxes for the images. We discuss the limitations and ethical considerations of our pipeline in the Appendix.

4.1 Synthetically Editing Images

Leveraging advancements in text-conditioned image generation and editing, we use an instruction-based model, InstructPix2Pix [13], for editing objects in an image – referred to as the *source* image – while keeping the background unchanged. We edit source images from COCO that (i) contain only one person, inferred from the number of person bounding boxes; and (ii) have a defined gender label, as defined in Sec. 3.2. These restrictions remove ambiguity. Next, we crop the image to the single person bounding box and feed it to InstructPix2Pix [13] along with multiple edit instructions for each attribute label, e.g., "make this person more *masculine/feminine*". Refer to Tab. 6 for the complete set of instruction templates. The edited person is then replaced in the source image. By only editing the appearance of the person in the image, we preserve the background content and minimize distortion – empirically, we found editing the entire *source image* rather than just the *source person* produced lower quality edits with significant hallucination. For further implementation details, refer to the Appendix.

4.2 Automatic Quality Filtering of Edited Images

The synthetic edits with InstructPix2Pix [13] can often be of low quality or fail to edit the source person's attribute into the target attribute. In order to ensure the quality and gender accuracy of our synthetic image sets, we introduce an automatic filtering method using K-Nearest Neighbor (KNN), similar to [57] who use KNN to score GAN-generated images.

First, we embed a collection of (i) source person bounding boxes, denoted as $R = \{r_1, r_2, ..., r_n\}$, and (ii) synthetically-edited person bounding boxes, denoted as $S = \{s_1, s_2, ..., s_m\}$ using CLIP. For each synthetic box s_i, we identify its K-nearest neighbors in this feature space, denoted as $N_{s_i} = \text{KNN}(s_i, R \cup S)$ using the Euclidean distance between the embeddings. If the proportion of real images within N_{s_i}, denoted as $P_R(s_i)$, and the proportion of images with the target gender of s_i, denoted as $P_G(s_i)$, exceed the thresholds τ_R and τ_G respectively, the edited image s_i is accepted:

$$P_R(s_i) = \frac{1}{K} \sum_{r \in N_{s_i}} \mathbb{1}(r \in R) \quad \text{and} \quad P_G(s_i) = \frac{1}{K} \sum_{r \in N_{s_i}} \mathbb{1}(\text{gender}(r) = \text{gender}(s_i)), \tag{1}$$

$$\text{accept}(s_i) = \begin{cases} 1 & \text{if } P_R(s_i) > \tau_R \text{ and } P_G(s_i) > \tau_G \\ 0 & \text{otherwise.} \end{cases} \tag{2}$$

This process ensures that the accepted images are of high quality and accurately reflect the target gender change. We only retain images where the entire set of edits per unique COCO ID has at least one accepted male and female edit, then randomly select one edit for each gender from images that pass the filter. For examples of edits at each decile of τ_R, see the Appendix. We verify the quality of the GENSYNTH dataset automatically and manually in the Appendix.

5 Benchmarking CLIP

5.1 Evaluation Setup

We use the following three datasets for evaluation: **GENSYNTH** consists of 7,946 images that have been generated and filtered as discussed in Sec. 4. It consists of 3,973 unique COCO images from the train set (62.6% of which were originally male), with a male and female edit for each. **COCO-♀🚶** consists of 3,973 original (unedited) images with the same unique COCO IDs as GENSYNTH. All

Table 2: Comparison of Gender Bias between CLIP-like models on COCO-Person datasets. We report the MaxSkew@K in caption-to-image retrieval of gender-neutralised captions. We compare CLIP [5] and CLIP-clip [20]. We also report zero-shot image classification accuracy on ImageNet1K [58].

COCO-Person Dataset	Model	Gender Bias ↓		ImageNet1k Acc. (%) ↑
		MaxSkew@25	MaxSkew@100	
COCO$_{♀}$	CLIP	0.27	0.20	63.2
	CLIP-clip$_{m=100}$	0.23	0.16	60.1
COCO$_{♀_{\text{Bal}}}$	CLIP	0.26	0.20	63.2
	CLIP-clip$_{m=100}$	0.22	0.15	60.1
GENSYNTH	CLIP	0.23	0.18	63.2
	CLIP-clip$_{m=100}$	0.22	0.17	60.1

images contain a single person, whose gender can be identified from the caption. **COCO$_{♀_{\text{Bal}}}$** consists of 2,970 unique images from COCO$_{♀}$, randomly sampled such that there is an equal number of male and female images. We use 5 different random seeds and report average results.

We compute MaxSkew@K for CLIP [5] and CLIP-clip [20], with $m = 100$ clipped dimensions computed on COCO train 2017. We use the ViT-B/32 variant for both models. Refer to Appendix for evaluation with other debiased CLIP-like models. We only report MaxSkew@K, as we showed in Sec. 3.3 that Bias@K is not a suitable measure of model bias due to its signed nature – where for balanced data *with* spurious correlations it shows close to zero bias due to biases in both directions canceling each other out.

5.2 Results

In Tab. 2 we compare the gender bias of the CLIP models for the three datasets defined in Sec. 5.1. We find debiased CLIP (CLIP-clip$_{m=100}$) records substantially lower bias on both unbalanced data (COCO$_{♀}$) and balanced (COCO$_{♀_{\text{Bal}}}$) data. This is because, as we noted in Sec. 3.3, balanced but biased data still contains spurious correlations related to gender. However, we observe that both models show very similar bias on the balanced and debiased [3] GENSYNTH data. This almost zero difference in bias is a result in itself – it means that the positive debiasing result on COCO$_{♀_{\text{Bal}}}$ and COCO$_{♀}$ are due to dataset bias. Overall, these findings suggest that intrinsic dataset bias, specifically spurious correlations, is artificially skewing the interpretations and comparisons of model bias. This reinforces the need for balanced data with no spurious correlations and shows the utility of our proposed pipeline for dataset debiasing.

6 Conclusion

The reliability of reported *model biases* in VLMs is affected by the interaction between *dataset bias* and choice of bias metric. In this paper, we demonstrated that naturalistic images from COCO have spurious correlations in image context with gender, which in turn affects how much trust can be placed in commonly-used metrics such as Bias@K: when measuring *model bias*, we may in fact be measuring *dataset bias*. To mitigate these problems, we proposed a pipeline for editing open-domain images at scale, creating gender-balanced contrast sets where the semantic content of the image remains the same except the person bounding box. Our method does not require manual auditing or image curation, relying instead on an effective automatic filtering method. Using this synthetically-created contrast set (GENSYNTH) we found that state-of-the-art CLIP-like models measure similarly on gender bias suggesting that measurements of model gender bias can largely be attributed to spurious model associations with gender (such as scene or background information) rather than gender itself. Through these subsequent angles of investigation, we conclude that only focusing on model bias while ignoring how dataset artefacts affect bias metrics paints an unreliable picture of identity-based bias in VLMs. We hope our work contributes to an ongoing discussion of how to seek improved representation and diversity of identity groups in image-captioning datasets, both now and in the future.

[3] Balanced in respect to gender ratio and debiased in respect to spurious correlations.

References

[1] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International Conference on Machine Learning*, pages 8821–8831. PMLR, 2021.

[2] MidJourney. Home Page, May 2023.

[3] Rob Salkowitz. Midjourney Founder David Holz On The Impact Of AI On Art, Imagination And The Creative Economy. *Forbes*, September 2022.

[4] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.

[5] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International Conference on Machine Learning*, pages 8748–8763. PMLR, 2021.

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

[7] 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. *arXiv preprint arXiv:2212.07143*, 2022.

[8] 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.

[9] Hugo Berg, Siobhan Mackenzie Hall, Yash Bhalgat, Wonsuk Yang, Hannah Rose Kirk, Aleksandar Shtedritski, and Max Bain. A prompt array keeps the bias away: Debiasing vision-language models with adversarial learning. *arXiv preprint arXiv:2203.11933*, 2022.

[10] 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.

[11] Kimmo Kärkkäinen and Jungseock Joo. Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2021.

[12] Matt Gardner, Yoav Artzi, Victoria Basmova, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, et al. Evaluating models' local decision boundaries via contrast sets. *arXiv preprint arXiv:2004.02709*, 2020.

[13] Tim Brooks, Aleksander Holynski, and Alexei A Efros. Instructpix2pix: Learning to follow image editing instructions. *arXiv preprint arXiv:2211.09800*, 2022.

[14] Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollár, and C Lawrence Zitnick. Microsoft coco captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325*, 2015.

[15] Mengmeng Wang, Jiazheng Xing, and Yong Liu. Actionclip: A new paradigm for video action recognition. *arXiv preprint arXiv:2109.08472*, 2021.

[16] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM Computing Surveys (CSUR)*, 54(6):1–35, 2021.

[17] Sahil Verma and Julia Rubin. Fairness definitions explained. pages 1–7. IEEE Computer Society, 5 2018.

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

[19] Felix Friedrich, Patrick Schramowski, Manuel Brack, Lukas Struppek, Dominik Hintersdorf, Sasha Luccioni, and Kristian Kersting. Fair diffusion: Instructing text-to-image generation models on fairness. *arXiv preprint arXiv:2302.10893*, 2023.

[20] Jialu Wang, Yang Liu, and Xin Wang. Are gender-neutral queries really gender-neutral? mitigating gender bias in image search. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 1995–2008, 2021.

[21] 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*, pages 2221–2231, 2019.

[22] Ashish Seth, Mayur Hemani, and Chirag Agarwal. Dear: Debiasing vision-language models with additive residuals. *arXiv preprint arXiv:2303.10431*, 2023.

[23] Junyang Wang, Yi Zhang, and Jitao Sang. Fairclip: Social bias elimination based on attribute prototype learning and representation neutralization. *arXiv preprint arXiv:2210.14562*, 2022.

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

[25] Joy Buolamwini and Timnit Gebru. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR, 2018.

[26] Tianlu Wang, Jieyu Zhao, Mark Yatskar, Kai-Wei Chang, and Vicente Ordonez. Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 5310–5319, 2019.

[27] Terrance De Vries, Ishan Misra, Changhan Wang, and Laurens Van der Maaten. Does object recognition work for everyone? In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 52–59, 2019.

[28] Antonio Torralba and Alexei A Efros. Unbiased look at dataset bias. In *CVPR 2011*, pages 1521–1528. IEEE, 2011.

[29] Angelina Wang and Olga Russakovsky. Overcoming bias in pretrained models by manipulating the finetuning dataset. *arXiv preprint arXiv:2303.06167*, 2023.

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

[31] Carsten Schwemmer, Carly Knight, Emily D Bello-Pardo, Stan Oklobdzija, Martijn Schoonvelde, and Jeffrey W Lockhart. Diagnosing gender bias in image recognition systems. *Socius*, 6:2378023120967171, 2020.

[32] Emiel van Miltenburg. Stereotyping and bias in the flickr30k dataset. In *Proceedings of Multimodal Corpora: Computer vision and language processing (MMC 2016)*, pages 1–4. 2016.

[33] Nicole Meister, Dora Zhao, Angelina Wang, Vikram V Ramaswamy, Ruth Fong, and Olga Russakovsky. Gender artifacts in visual datasets. *arXiv preprint arXiv:2206.09191*, 2022.

[34] 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. *arXiv preprint arXiv:2210.08402*, 2022.

[35] Kaiyu Yang, Klint Qinami, Li Fei-Fei, Jia Deng, and Olga Russakovsky. Towards fairer datasets: Filtering and balancing the distribution of the people subtree in the imagenet hierarchy. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pages 547–558, 2020.

[36] Yuki M Asano, Christian Rupprecht, Andrew Zisserman, and Andrea Vedaldi. Pass: An imagenet replacement for self-supervised pretraining without humans. 2021.

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

[38] Umberto Michieli, Matteo Biasetton, Gianluca Agresti, and Pietro Zanuttigh. Adversarial learning and self-teaching techniques for domain adaptation in semantic segmentation. *IEEE Transactions on Intelligent Vehicles*, 5(3):508–518, 2020.

[39] Alexey Dosovitskiy, Philipp Fischer, Eddy Ilg, Philip Hausser, Caner Hazirbas, Vladimir Golkov, Patrick Van Der Smagt, Daniel Cremers, and Thomas Brox. Flownet: Learning optical flow with convolutional networks. pages 2758–2766, 2015.

[40] Shuran Song, Fisher Yu, Andy Zeng, Angel X Chang, Manolis Savva, and Thomas Funkhouser. Semantic scene completion from a single depth image. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 1746–1754, 2017.

[41] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in Neural Information Processing Systems*, 35:36479–36494, 2022.

[42] William Peebles, Jun-Yan Zhu, Richard Zhang, Antonio Torralba, Alexei A Efros, and Eli Shechtman. Gan-supervised dense visual alignment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 13470–13481, 2022.

[43] Daiqing Li, Huan Ling, Seung Wook Kim, Karsten Kreis, Sanja Fidler, and Antonio Torralba. Bigdataset-gan: Synthesizing imagenet with pixel-wise annotations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 21330–21340, 2022.

[44] Andrew Zhai and Hao-Yu Wu. Classification is a strong baseline for deep metric learning. *arXiv preprint arXiv:1811.12649*, 2018.

[45] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626*, 2022.

[46] Narek Tumanyan, Michal Geyer, Shai Bagon, and Tali Dekel. Plug-and-play diffusion features for text-driven image-to-image translation. *arXiv preprint arXiv:2211.12572*, 2022.

[47] Prasanna Sattigeri, Samuel C Hoffman, Vijil Chenthamarakshan, and Kush R Varshney. Fairness gan: Generating datasets with fairness properties using a generative adversarial network. *IBM Journal of Research and Development*, 63(4/5):3–1, 2019.

[48] Erik Altman. Synthesizing credit card transactions. In *Proceedings of the Second ACM International Conference on AI in Finance*, pages 1–9, 2021.

[49] Brian Belgodere, Pierre Dognin, Adam Ivankay, Igor Melnyk, Youssef Mroueh, Aleksandra Mojsilovic, Jiri Navartil, Apoorva Nitsure, Inkit Padhi, Mattia Rigotti, et al. Auditing and generating synthetic data with controllable trust trade-offs. *arXiv preprint arXiv:2304.10819*, 2023.

[50] Karan Bhanot, Miao Qi, John S Erickson, Isabelle Guyon, and Kristin P Bennett. The problem of fairness in synthetic healthcare data. *Entropy*, 23(9):1165, 2021.

[51] Yingzhou Lu, Huazheng Wang, and Wenqi Wei. Machine learning for synthetic data generation: a review. *arXiv preprint arXiv:2302.04062*, 2023.

[52] Isabal Hermes. Gender representation in ai – part 1: Utilizing stylegan to explore gender directions in face image editing, 8 2022.

[53] Vikram V Ramaswamy, Sunnie SY Kim, and Olga Russakovsky. Fair attribute classification through latent space de-biasing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9301–9310, 2021.

[54] Emily Denton, Ben Hutchinson, Margaret Mitchell, Timnit Gebru, and Andrew Zaldivar. Image counterfactual sensitivity analysis for detecting unintended bias. *arXiv preprint arXiv:1906.06439*, 2019.

[55] Rebecca Qian, Candace Ross, Jude Fernandes, Eric Smith, Douwe Kiela, and Adina Williams. Perturbation augmentation for fairer nlp. *arXiv preprint arXiv:2205.12586*, 2022.

[56] 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 *Computer Vision–ECCV 2014: 13th European Conference, Zurich, Switzerland, September 6-12, 2014, Proceedings, Part V 13*, pages 740–755. Springer, 2014.

[57] Shuyang Gu, Jianmin Bao, Dong Chen, and Fang Wen. Giqa: Generated image quality assessment. In *Computer Vision–ECCV 2020: 16th European Conference, Glasgow, UK, August 23–28, 2020, Proceedings, Part XI 16*, pages 369–385. Springer, 2020.

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

[59] Su Lin Blodgett, Gilsinia Lopez, Alexandra Olteanu, Robert Sim, and Hanna Wallach. Stereotyping norwegian salmon: An inventory of pitfalls in fairness benchmark datasets. pages 1004–1015. Association for Computational Linguistics (ACL), 2021.

[60] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE signal processing letters*, 23(10):1499–1503, 2016.

[61] 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. *Commun. ACM*, 59(2):64–73, 2016.

[62] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

[63] Gregory Plumb, Marco Tulio Ribeiro, and Ameet Talwalkar. Finding and fixing spurious patterns with explanations. *arXiv preprint arXiv:2106.02112*, 2021.

[64] Gabriel Ilharco, Mitchell Wortsman, Ross Wightman, Cade Gordon, Nicholas Carlini, Rohan Taori, Achal Dave, Vaishaal Shankar, Hongseok Namkoong, John Miller, Hannaneh Hajishirzi, Ali Farhadi, and Ludwig Schmidt. Openclip, July 2021.

[65] Mathilde Caron, Hugo Touvron, Ishan Misra, Hervé Jégou, Julien Mairal, Piotr Bojanowski, and Armand Joulin. Emerging properties in self-supervised vision transformers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 9650–9660, 2021.

Appendix

A Limitations and Ethical Considerations

Synthetic Shifts. By generating synthetic data, we are creating a new evaluation distribution that does not necessarily represent the real-world distribution of the respective categories. This distribution shift can also be forced in contexts where it does not necessarily make sense to either face swap or make gender edits due to factual histories or biological identity [59].

Assumptions of Binary Gender. Our data relies on the binary gender labels from the COCO and FairFace datasets and is necessarily influenced by our respective identities. COCO also presents limitations regarding race, ethnicity, and other sensitive attributes. We acknowledge this approach of using binary gender and making reference to perceived gender based on appearance oversimplifies the complexity of gender identity and biological sex, and risks erasing representation of non-binary people. Despite attempts to mitigate this limitation using terms such as "masculine" and "feminine", the resulting edits were often unusable (due to existing biases in generative models), necessitating reliance on binary and narrow terms. We advocate for future work that encodes and represents non-binary gender in datasets, and improves generalisation in generative models to non-binary terms.

Stacking Biases. Our pipeline may inadvertently introduce biases from the generative model via stereotypical representations of perceived gender, e.g., if "make this person more feminine" over-emphasises pink clothes, or "make this person more masculine" over-emphasises beards. The automatic filtering step also tends to favour images with simple scene arrangements. Some generated images were identified as NSFW, a consequence of training on large-scale internet datasets [6]. Future work could integrate into our pipeline more capable and fair generative models.

B Implementation Details

Here we provide additional implementation details about our method.

B.1 Gendered Words and Caption Editing

In Tab. 3 we show the gendered words (Masculine, Feminine) that we use for assigning each caption a gender label. Captions without either a masculine or feminine word, or captions with matches from both of these lists are labeled as *undefined*. For switching or neutralising the gender in a caption, we map words across the rows of Tab. 3, so for example "she" could be replaced with "he" or "they". In Tab. 4 we show sentences that have been gender-neutralised.

Table 3: **Gendered word pairs.** We the Masculine and Feminine words in order to classify the gender of a person in an image given its caption. When editing the gender of a caption or making it gender-neutral, we use the word from the corresponding pair for the opposite gender or the gender-neutral word, respectively.

Masculine	Feminine	Neutral
man	woman	person
men	women	people
male	female	person
boy	girl	child
boys	girls	children
gentleman	lady	person
father	mother	parent
husband	wife	partner
boyfriend	girlfriend	partner
brother	sister	sibling
son	daughter	child
he	she	they
his	hers	their
him	her	them

Table 4: **Examples of gender-neutralised captions.** We show example original COCO captions with their gender-neutralised replacements, using the corresponding words from Tab. 3

Original	Neutral
The **woman** brushes **her** teeth in the bathroom.	The **person** brushes **their** teeth in the bathroom.
A **man** sleeping with **his** cat next to **him**.	A **person** sleeping with **their** car next to **them**.
Two **women** and two **girls** in makeup and one is talking on a cellphone.	Two **people** and two **children** in makeup and one is talking on a cellphone.

B.2 Image editing

Here we provide additional details on the two image editing pipelines in the paper – our proposed method GENSYNTH, and the weak baseline GENSWAP.

GENSYNTH We edit the COCO train set images by applying InstructPix2Pix [13] on person crops (bounding boxes) with gender-editing instructions, as described in the main paper. We run InstructPix2Pix for 500 denoising steps, and for each instruction, we generate an image with two text guiding scales: 9.5 and 15. We found that a smaller guiding scale sometimes does not produce the required edit, whereas too large a scale results in an image that does not look natural. Using both scales ensures there are multiple candidates for the edited image, and then we can use the filtering pipeline to discard bad edits.

GENSWAP We use the MTCNN face detector [60] to detect faces in the COCO images (for the same subset in GENSYNTH), and replace them with faces from the FairFace repository [11]. FairFace is a collection of face crops from the YFCC-100M dataset [61], labeled with gender, race and age. We only use images whose age attribute is greater than 19 and randomly sample a face crop from the target gender.

B.3 Filtering

For the KNN filter, we set the neighbourhood size $K = 50$, and the thresholds $\tau_R = 0.08$ and $\tau_G = 0.5$.

C Bias Measures

Bias@K [20] measures the proportions of masculine and feminine images in the retrievals of a search result with a gender-neutral text query. For an image I, we define a function $g(I) = \text{male}$ if there are only individuals who appear as men in the image, and $g(I) = \text{female}$ if there are only individuals who appear as women. Given a set of K retrieved images $\mathcal{R}_K(q)$ for a query q, we count the images of apparent men and women as:

$$N_{\text{male}} = \sum_{I \in \mathcal{R}_K(q)} \mathbb{1}[g(I) = \text{male}] \quad \text{and} \quad N_{\text{female}} = \sum_{I \in \mathcal{R}_K(q)} \mathbb{1}[g(I) = \text{female}].$$

We define the gender bias metric as:

$$\delta_K(q) = \begin{cases} 0, & N_{\text{male}} + N_{\text{female}} = 0 \\ \frac{N_{\text{male}} - N_{\text{female}}}{N_{\text{male}} + N_{\text{female}}}, & \text{otherwise.} \end{cases}.$$

For a whole query set Q, we define:

$$\text{Bias@K} = \frac{1}{|Q|} \sum_{q \in Q} \delta_K(q). \tag{3}$$

Skew@K [9, 21] measures the difference between the desired proportion of image attributes in $\mathcal{R}_k(q)$ for the query q and the actual proportion. Let the desired proportion of images with attribute label A in the set of retrieved images be $p_{d,q,A} \in [0, 1]$ and the actual proportion be $p_{\mathcal{R}(q),q,A} \in [0, 1]$. The Skew@K of $\mathcal{R}(q)$ for an attribute label $A \in \mathcal{A}$ is:

$$\text{Skew@K}(\mathcal{R}(q)) = \ln \frac{p_{\mathcal{R}_K(q),q,A}}{p_{d,q,A}}, \tag{4}$$

where the desired proportion $p_{d,q,A}$ is the actual attribute distribution over the entire dataset. A disadvantage of Skew@K is that it only measures bias with respect to a single attribute at a time and must be aggregated to give a holistic view of the bias over all attributes. We follow [9] and take the maximum Skew@K among all attribute labels A of the images for a given text query q:

$$\text{MaxSkew@K}(\mathcal{R}(q)) = \max_{A_i \in \mathcal{A}} \text{Skew}_{A_i}\text{@K}(\mathcal{R}(q)), \tag{5}$$

which gives us the "largest unfair advantage" [21] belonging to images within a given attribute. In our work, a MaxSkew@K of 0 for the attribute gender and a given text query q implies that men and women are equally represented in the retrieved set of K images $\mathcal{R}_K(q)$. We ignore all images with undefined attribute labels (in this case gender) when measuring MaxSkew@K.

D Spurious Correlations Analysis

COCO contains more than twice as many male images as it does female ones. This will inevitably affect retrieval-based bias metrics, as there will be more male images in the retrievals. One naïve way to fix this is to undersample the male images in order to arrive at a *Balanced* COCO dataset. However, ensuring equal distribution of demographic attributes does not necessarily ensure the dataset is unbiased as a whole. Spurious correlations can result in subsets of the data being highly correlated with certain attributes. Works such as [29, 33] exemplify the spurious correlations in COCO. We further extend these by showing the existence of spurious correlations in the COCO Captions dataset. We take two approaches to evidence these spurious correlations.

D.1 K-means Clusters with Caption Embeddings

First, we find semantic clusters of captions and evaluate the gender balance within them. For every image I_n, we embed its gender-neutralised captions C_n^k, where $k = \{1, \dots, K\}$ represents the K captions of the image, with RoBERTa [62] to get features f_n^k. We average the features to get $f_n = \frac{1}{K}\sum_{k=1}^{K} f_n^k$. Next, we cluster the features $f_n, n = \{1, \dots, N\}$ into $M = 20$ clusters with K-Means. Finally, for each cluster, we extract salient words using Latent Dirichlet Allocation (LDA) and give a manually-defined cluster label. In Fig. 2 we show a t-SNE representation of the discovered clusters, together with the degree of male over-representation. We see that in sports-related concepts men are over-represented, whereas in scenes in kitchens, bathrooms, streets, and parks, women are over-represented. For a comparison of this analysis to a keyword-based analysis, and a list of all discovered classes and salient words according to LDA, please see appendix D.2 and appendix D.3.

D.2 Using Discovered Clusters vs COCO Classes

While prior works such as [63] use co-appearance of COCO classes, e.g. "tennis racket" and "person" to explore spurious correlations in COCO, we opt for discovering such keywords automatically from captions. We do so for two reasons. Firstly, using class co-occurrence simplifies the spurious correlations that exist in the dataset. For example, take the discovered clusters for $\texttt{leisure}_{\text{street}}$ and $\texttt{sports}_{\text{cycling,motorcycling}}$. Both appear on the street and considering co-occurence of COCO classes such as "car", "motorcycle", "water hydrant" could group the two clusters together. These two clusters exhibit opposite biases and if grouped together, would result in a close to zero overall bias. In contrast, captions refer to the activity the subject of the caption is performing, allowing us to separate semantically different activities. Secondly, our analysis only requires image captions, which are cheaper to obtain than object labels, and might be more generalizable to other datasets.

D.3 Discovered Clusters

In Tab. 5 we show the 20 discovered clusters using K-Means, together with the top 10 salient words according to LDA. For each cluster, we show the male-overrepresentation factor, i.e., the difference between the percentage of images in that particular cluster relative to the percentage of male images in the person class of COCO as a whole.

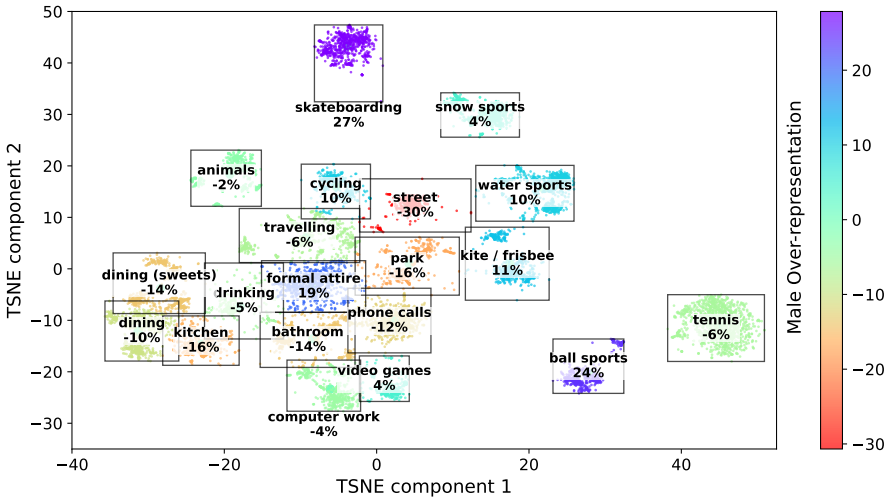


Figure 2: t-SNE clusters ($M = 20$) of gender-neutralised caption embeddings. Each cluster is manually assigned a name, then coloured and labelled according to its male over-representation factor. The male over-representation factor is the difference between the percentage of male images in the particular cluster and the percentage of male images overall in the dataset.

Table 5: **Discovered clusters in COCO Captions.** We show all 20 clusters with their manually assigned names, together with the top 10 words according to LDA. ΔM represents the deviation from gender parity for males.

Name	**Words**	Δ**M** (%)
$\text{dining}_{\text{drinking}}$	wine, glass, holding, scissors, table, sitting, bottle, drinking, pouring, standing	-5.7
$\text{dining}_{\text{sweets}}$	cake, banana, donut, doughnut, holding, eating, candle, table, sitting, birthday	-14.0
$\text{dining}_{\text{mains}}$	pizza, eating, table, food, sandwich, sitting, holding, slice, hot, dog	-10.3
$\text{sports}_{\text{tennis}}$	tennis, court, racket, ball, player, racquet, hit, holding, swinging, playing	-6.0
$\text{sports}_{\text{snow}}$	ski, snow, slope, skiing, skier, snowboard, snowy, snowboarder, standing, hill	4.7
$\text{sports}_{\text{skateboarding}}$	skateboard, skate, skateboarder, riding, trick, skateboarding, ramp, young, board, child	27.9
$\text{sports}_{\text{ball}}$	baseball, bat, player, ball, soccer, field, pitch, holding, game, pitcher	24.0
$\text{sports}_{\text{kite,frisbee}}$	frisbee, kite, playing, holding, field, beach, throwing, flying, standing, child	11.6
$\text{sports}_{\text{surfing}}$	surfboard, wave, surf, surfer, riding, water, surfing, board, ocean, beach	10.1
$\text{sports}_{\text{cycling,motorcycling}}$	motorcycle, riding, bike, bicycle, street, sitting, next, standing, ride, motor	10.5
$\text{leisure}_{\text{street}}$	umbrella, holding, hydrant, standing, rain, fire, walking, street, child, black	-30.7
$\text{leisure}_{\text{park}}$	sitting, dog, bench, next, holding, park, child, two, sits, frisbee	-16.9
formal attire	tie, wearing, suit, standing, shirt, glass, shirt, black, white, young	19.7
computer work	laptop, sitting, computer, bed, couch, desk, room, table, using, front	-4.6
animals	horse, elephant, giraffe, riding, cow, standing, sheep, next, two, brown	-2.9
video games	wii, game, remote, controller, playing, video, Nintendo, holding, room, standing	4.8
kitchen	kitchen, food, standing, refrigerator, oven, cooking, counter, chef, preparing, holding	-16.2
bathroom	brushing, mirror, teeth, bathroom, cat, toothbrush, taking, toilet, holding, child	-14.0
travelling	standing, bear, teddy, luggage, train, next, street, bus, holding, suitcase	-6.7
phone calls	phone, cell, talking, holding, sitting, cellphone, standing, looking, wearing, young	-12.8

D.4 Spurious Correlations Classifier

Following [31], we investigate the presence of spurious correlations by training classifiers to predict binary gender labels of images and captions where the explicit gender information is removed for both training and testing. Specifically, for the image classifier (ResNet-50) we replace all person bounding boxes with black pixels; and for the caption classifier (BERT-base) we use the gender-neutralised captions. The training and testing data is COCO train and validation defined in Sec. 3.2, but with undefined images dropped. On unseen data, the text-only classifier on gender-neutralised captions achieves 78.0% AUC and the image-only classifier on person-masked images achieves 63.4% AUC. Given that a random chance model achieves 50% AUC and an image classifier on unmasked images achieves 71.9% AUC, it is clear that spurious correlations in the image, as well as biases in the caption, provide a significant signal to predict gender of the person in the image even when there is no explicit gender information.

E Prompt Editing Templates

Tab. 6 contains the complete set of edit instructions input to InstructPix2Pix to edit the single person bounding box for each attribute label.

Table 6: Templates used for prompt editing.

Template	Instruction	
	Feminine	**Masculine**
Make this person more {}	feminine	masculine
Make this person look like a {}	woman	man
Turn this person into a {}	woman	man
Convert this into a {}	woman	man

F Verifying the Quality of GENSYNTH

We evaluate the quality of the GENSYNTH dataset in three ways. First, we perform the human evaluation study described in appendix F.1. The results affirm our pipeline's successful gender-targeted editing, with high human agreement. Second, we automatically measure the correctness of the targeted gender edit, by using CLIP to zero-shot classify the gender of people in the images. Third, to evaluate the semantic similarity of the edited image to the caption, we measure the text-to-image retrieval performance of CLIP on the synthetic text-image captions. For this, we edit the captions using the reverse procedure in Sec. 3.2 to reflect the gender of the person in the edited image. Then, for each image I_i in GENSYNTH, where $i \in \{1, 2, \ldots, N\}$, we have a set of n captions C_i^j, $j \in \{1, 2, \ldots, n\}$. For each caption C_i^j, we perform a retrieval operation from the COCO validation set combined with the query image I_i, to find a set of K images that most closely match the caption, according to Euclidean distance of CLIP features. We denote this retrieved set as $R_i^j(K)$. The retrieval performance is evaluated using Recall at K (R@K), which is defined as $R@K = \frac{1}{Nn} \sum_{i=1}^{N} \sum_{j=1}^{n} \mathbb{1}(I_i \in R_i^j(K))$.

We compare GENSYNTH, against (i) the original COCO 2017 dataset (train set) of natural images containing persons; and (ii) a weak gender-editing baseline – GENSWAP. This baseline has the same unique COCO images as in GENSYNTH, but only with edited faces – we replace the detected face in the COCO image with a random face of the target gender from the FairFace dataset [11]. Additional implementations of GENSWAP are provided in appendix B.2.

As shown in Tab. 7, GENSYNTH leads to very similar zero-shot classification and retrieval results to the original COCO images. The filtering step significantly improves both metrics, successfully removing bad edits. The weak baseline, GENSWAP, consistently scores low, showing the importance of an effective editing method.

Table 7: Dataset comparison between the original COCO dataset of natural person images and synthetically edited COCO from the GENSWAP and GENSYNTH pipelines. We report the presence of Spurious Background (BG) Correlations, Zero-Shot (ZS) Gender Accuracy, and Text-to-Image Retrieval Recall@K (R@K) amongst COCO Val 5k images using CLIP. *Unfilt.* refers to the synthetic pipeline without automatic quality filtering.

COCO-Person Dataset	# Images	Edits per Image	Spurious BG. Correlations	ZS Gender Acc. (%) ↑	Text-to-Image Retrieval ↑		
					R@1	R@5	R@10
Original	11,541	-	✓	93.6	30.9	54.4	64.9
GENSWAP	3,973	2	✗	67.9	19.0	39.8	50.4
GENSYNTH (unfilt.)	11,541	16	✗	83.9	22.4	43.4	53.8
GENSYNTH	3,973	2	✗	95.5	29.2	52.8	62.8

F.1 Human Evaluation Study

Each of two annotators annotated the perceived gender of 100 images from the GENSYNTH dataset. They then annotated the perceived gender of the 100 original COCO images corresponding to the same IDs. The 100 GENSYNTH images were randomly sampled from the dataset without replacement so there were no repeats and no overlap between annotators. For the first annotator, their given labels matched the GENSYNTH gender label in 99% of images (99 images), and their given label matched the COCO original gender label in 95% of images. For the second annotator, there was 95% agreement in gender labels for the GENSYNTH images and 98% agreement in the COCO original images. In sum, these results show that our pipeline successfully edits the subject of the image to the target gender (e.g., from a man to a woman) as demonstrated by the high levels of human agreement.

G Extended Benchmarking of CLIP

Here we extend the analysis of CLIP models in the main paper. We evaluate the following models: (i) CLIP [5]; (ii) CLIP-clip [20], with $m = 100$ clipped dimensions computed on COCO train 2017; (iii) DebiasCLIP [9], which has been debiased on the FairFace dataset; and (iv) OpenCLIP [64] models trained on LAOIN 400M and 2BN datasets [34]. We use the ViT-B/32 variant for all models, except for DebiasCLIP, for which ViT-B/16 is used due to its availability from the authors.

In Tab. 8 we make a similar observation to the one discussed in the paper, where debiased models perform on par with other CLIP models on GENSYNTH.

Table 8: Comparison of Gender Bias between CLIP-like models on COCO-Person datasets. We report the MaxSkew@K in caption-to-image retrieval of gender-neutralised captions. We compare CLIP [5] and CLIP-clip [20], DebiasCLIP [9], and OpenCLIP [64] trained on LAOIN 400M & 2BN [34]. We additionally report zero-shot image classification accuracy on ImageNet1K [58].

COCO-Person Dataset	Model	Gender Bias ↓		ImageNet1k Acc. (%) ↑
		MaxSkew@25	MaxSkew@100	
COCO$_{♀}$	CLIP	0.27	0.20	63.2
	CLIP-clip$_{m=100}$	0.23	0.16	60.1
	DebiasCLIP	0.29	0.22	67.6
	OpenCLIP$_{400M}$	0.26	0.20	62.9
	OpenCLIP$_{2B}$	0.27	0.21	65.6
COCO$_{♀Bal}$	CLIP	$0.26_{\pm 0.00}$	$0.20_{\pm 0.00}$	63.2
	CLIP-clip$_{m=100}$	$0.22_{\pm 0.00}$	$0.15_{\pm 0.00}$	60.1
	DebiasCLIP	$0.28_{\pm 0.01}$	$0.21_{\pm 0.00}$	67.6
	OpenCLIP$_{400M}$	$0.27_{\pm 0.00}$	$0.20_{\pm 0.00}$	62.9
	OpenCLIP$_{2B}$	$0.27_{\pm 0.00}$	$0.21_{\pm 0.00}$	65.6
GENSYNTH	CLIP	0.23	0.18	63.2
	CLIP-clip$_{m=100}$	0.22	0.17	60.1
	DebiasCLIP	0.24	0.19	67.6
	OpenCLIP$_{400M}$	0.24	0.19	62.9
	OpenCLIP$_{2B}$	0.23	0.18	65.6

H Ablation Study

We ablate the use of a CLIP vision encoder in the KNN filtering pipeline. We replace it with a DINO ViT-B/16 [65] and repeat the analysis. We found that using DINO features is much more powerful when it comes to discriminating between the different images (real versus fake), and that the male and female images are better clustered. Accordingly, for the real vs. fake filter we use a neighborhood size of $K = 5{,}000$ and a threshold $\tau_R = 0.0002$ (i.e., the generated images have at least *one* real neighbour). For the male vs. female filter, we use a neighborhood size of $K = 50$ and a threshold $\tau_G = 0.4$. We end up with 571 unique COCO images, or 1,142 images in total (with a male and female edit for each unique image). The R@K results with this dataset are R@1 = 33.7%, R@5 = 57.1% and R@10 = 66.7%, and the zero-shot gender classification accuracy is 87.4%. Due to the different filtering, this dataset (with DINO filtering) is smaller than GENSYNTH and the results have higher variance, but are comparable to GENSYNTH.

We evaluate MaxSkew@K on this dataset in Tab. 9. We observe a similar trend to the GENSYNTH dataset, where bias results across models have a smaller variance than results on the unbalanced and balanced COCO$_{♀}^{♀}$ datasets. The absolute values of the bias metric are smaller, which we explain with the different images retrieved, and the variance that comes with that.

Table 9: Comparison of Gender Bias between CLIP-like models on the accepted images using DINO image embeddings for KNN filtering. We report the MaxSkew@K in caption-to-image retrieval of gender-neutralised captions. We compare CLIP [5], CLIP-clip [20]. We additionally report zero-shot image classification accuracy on ImageNet1K [58].

COCO-Person Dataset	Model	Gender Bias ↓		ImageNet1k Acc. (%) ↑
		MaxSkew@25	MaxSkew@100	
GENSYNTH (DINO)	CLIP	0.15	0.12	63.2
	CLIP-clip$_{m=100}$	0.13	0.10	60.1

I Qualitative Dataset Examples

In Fig. 3, we show gender edits for the GENSYNTH and GENSWAP datasets, alongside the original COCO image and ID. The GENSYNTH edits are more naturalistic than the GENSWAP edits, and also make changes to the body or clothing of the subject.

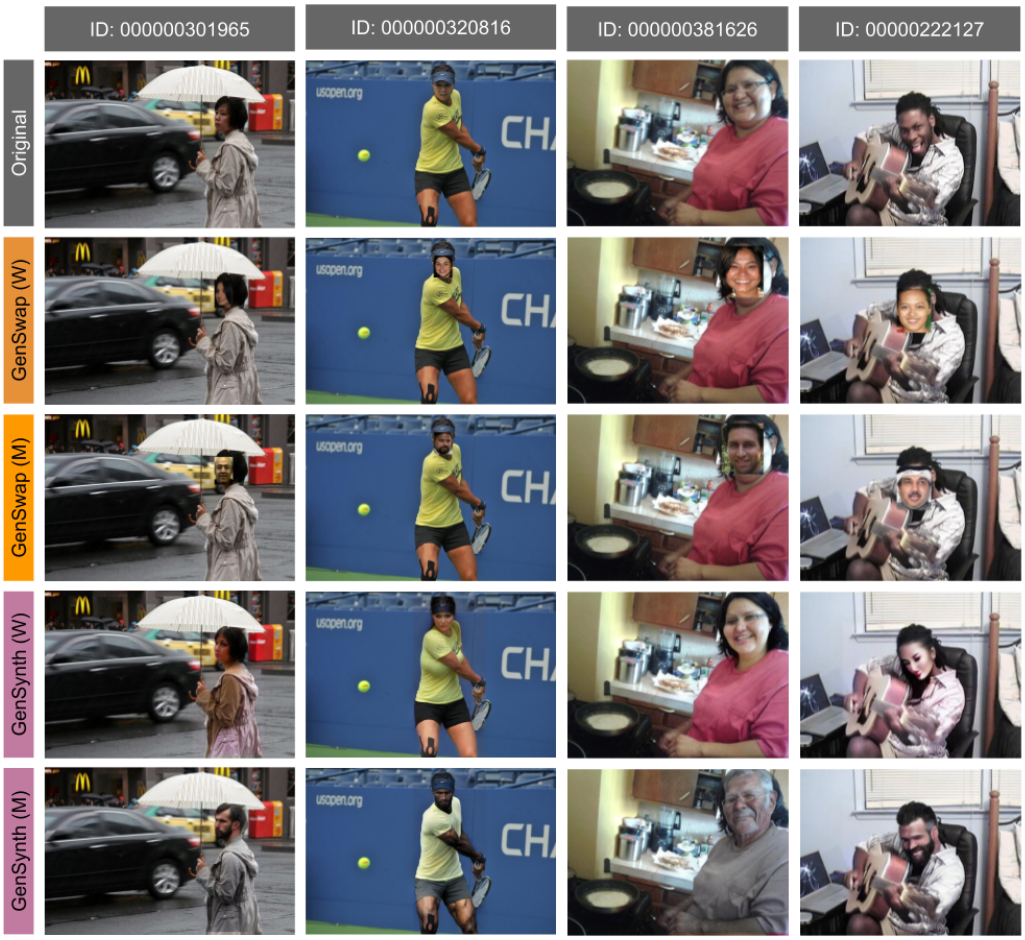
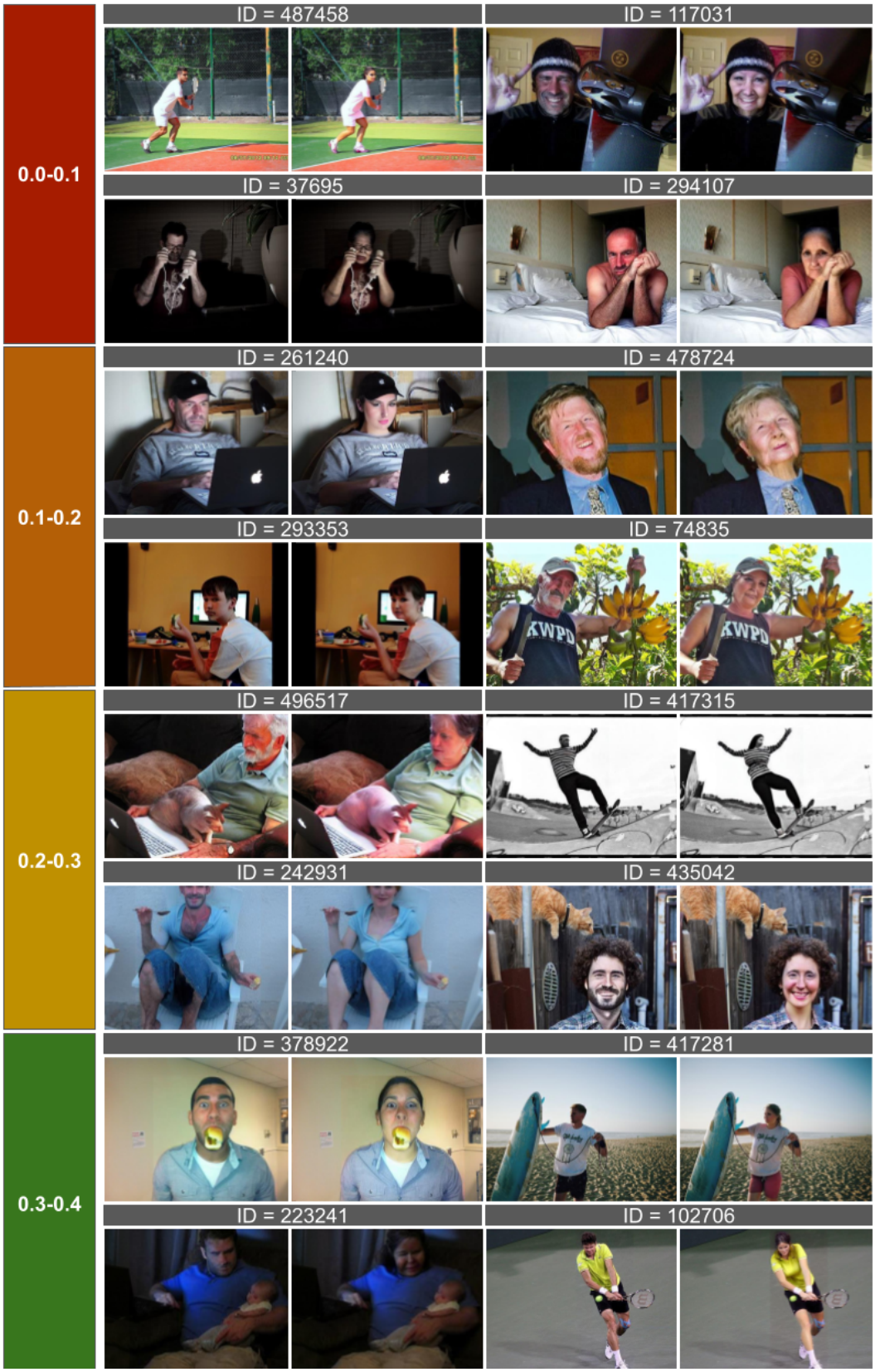


Figure 3: Randomly selected examples of GENSYNTH images showing a comparison to the original COCO image and the weak baseline GENSWAP.

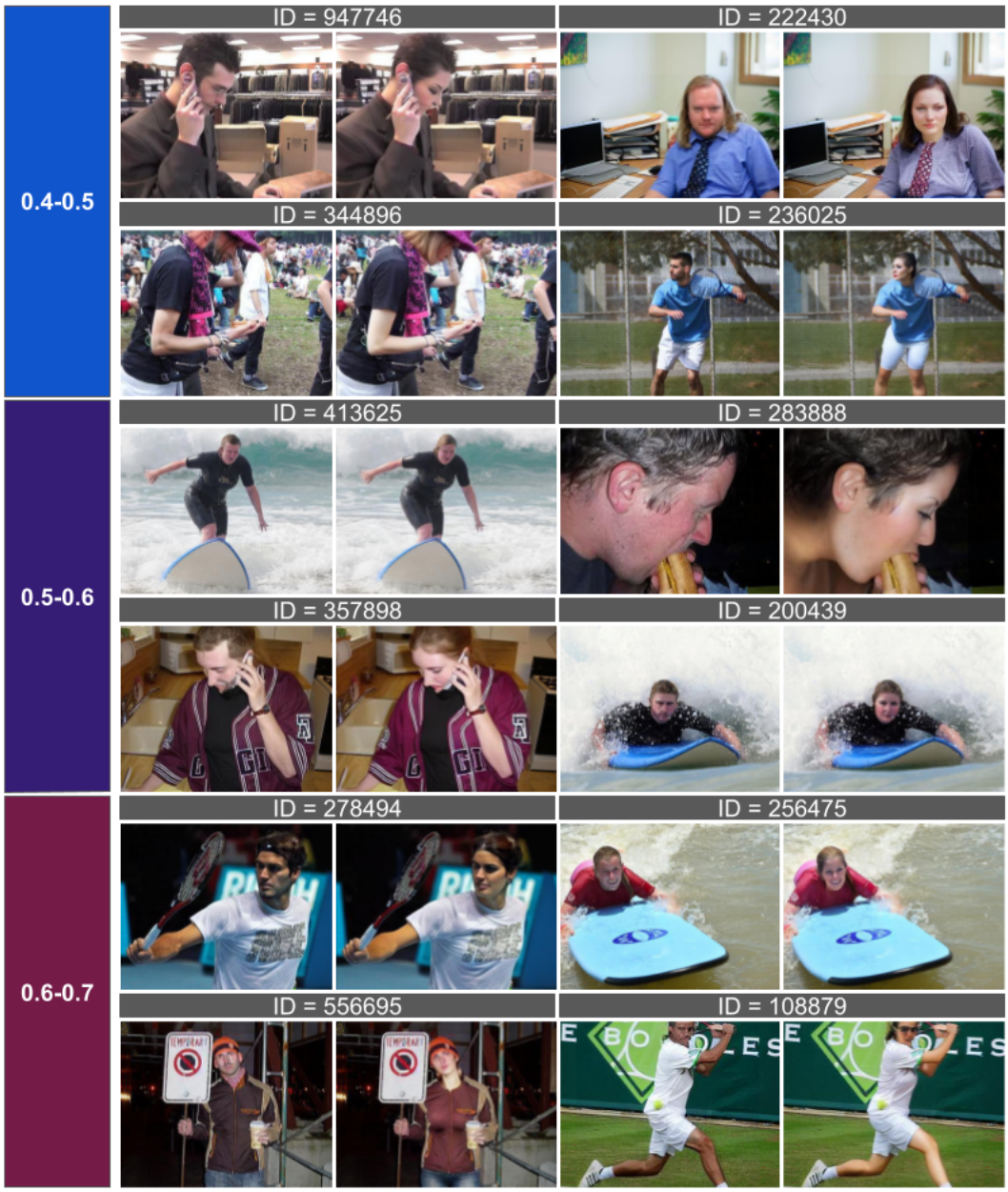
J Comparing Image Edits Across Filtering Thresholds

For each edited image, we calculate P_R, i.e., the ratio of real images versus fake images in the KNN clustering step. We then average P_R for each *pair* of images (the male and female edit). In Fig. 4 and appendix J, we show these randomly-selected pairs of gender edits from each decile of averaged P_R to demonstrate how our threshold filtering step improves the quality of the edited images.

Figure 4: Averaged KNN Score (P_R) for pairs of edited images using the GENSYNTH pipeline.



1st to 4th decile of scores.



5th to 8th decile of scores. Note that there was only one image with an averaged score between 0.7-0.8, and no images in the higher deciles.