Appendix

A Implementation Details

Here we provide additional implementation details about our method.

A.1 Gendered Words and Caption Editing

In Tab. 1 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. 1, so for example "she" could be replaced with "he" or "they". In Tab. 2 we show sentences that have been gender-neutralised.

Table 1: **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 2: **Examples of gender-neutralised captions.** We show example original COCO captions with their gender-neutralised replacements, using the corresponding words from Tab. 1

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.

A.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 Brooks et al. (2022) 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.

Table 3: **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

GENSWAP We use the MTCNN face detector Zhang et al. (2016) to detect faces in the COCO images (for the same subset in GENSYNTH), and replace them with faces from the FairFace repository Kärkkäinen and Joo (2021). FairFace is a collection of face crops from the YFCC-100M dataset Thomee et al. (2016), 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.

A.3 FILTERING

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

B SPURIOUS CORRELATIONS ANALYSIS

B.1 USING DISCOVERED CLUSTERS VS COCO CLASSES

While prior works such as Plumb et al. (2021) 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.

B.2 DISCOVERED CLUSTERS

In Tab. 3 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.

C Prompt Editing Templates

Tab. 4 containes the complete set of edit instructions input to InstructPix2Pix to edit the single person bounding box for each attribute label.

Table 4: 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

D 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.

E Extended Benchmarking of CLIP

Here we extend the analysis of CLIP models in the main paper. We evaluate the following models: (i) CLIP Radford et al. (2021); (ii) CLIP-clip Wang et al. (2021), with $m = 100$ clipped dimensions computed on COCO train 2017; (iii) DebiasCLIP Berg et al. (2022), which has been debiased on the FairFace dataset; and (iv) OpenCLIP Ilharco et al. (2021) models trained on LAOIN 400M and 2BN datasets Schuhmann et al. (2022). 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. 5 we make a similar observation to the one discussed in the paper, where debiased CLIP models perform on par with other CLIP models on GenSynth.

F 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 Caron et al. (2021) 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. 6. 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$_{\text{♀}}^{\text{♂}}$ 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 5: 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 Radford et al. (2021) and CLIP-clip Wang et al. (2021), DebiasCLIP Berg et al. (2022), and OpenCLIP Ilharco et al. (2021) trained on LAOIN 400M & 2BN Schuhmann et al. (2022). We additionally report zero-shot image classification accuracy on ImageNet1K Deng et al. (2009).

COCO-Person Dataset	Model	Gender Bias ↓		ImageNet1k Acc. (%) ↑
		MaxSkew@25	MaxSkew@100	
$\text{COCO}_{\female\male}$	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
$\text{COCO}_{\female\male\text{Bal}}$	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

Table 6: 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 Radford et al. (2021), CLIP-clip Wang et al. (2021). We additionally report zero-shot image classification accuracy on ImageNet1K Deng et al. (2009).

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

G QUALITATIVE DATASET EXAMPLES

In Fig. 1, 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.

H 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. 2, 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.

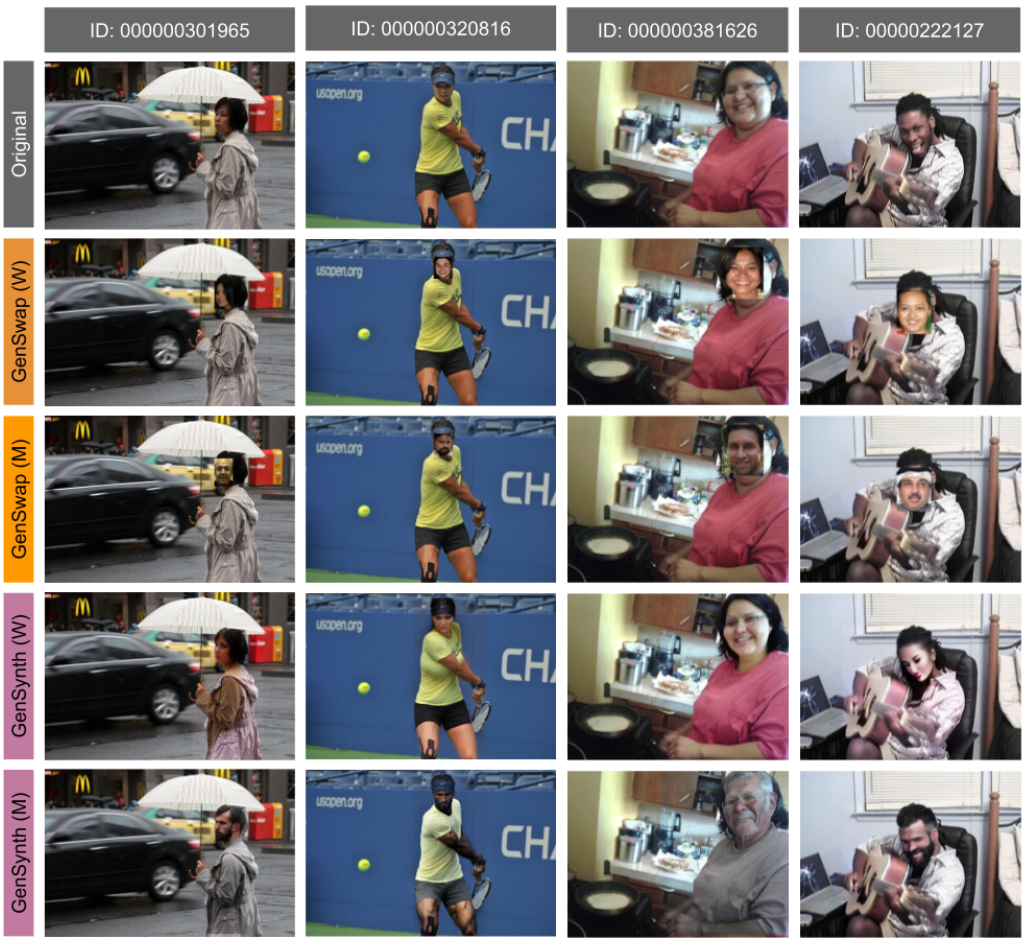
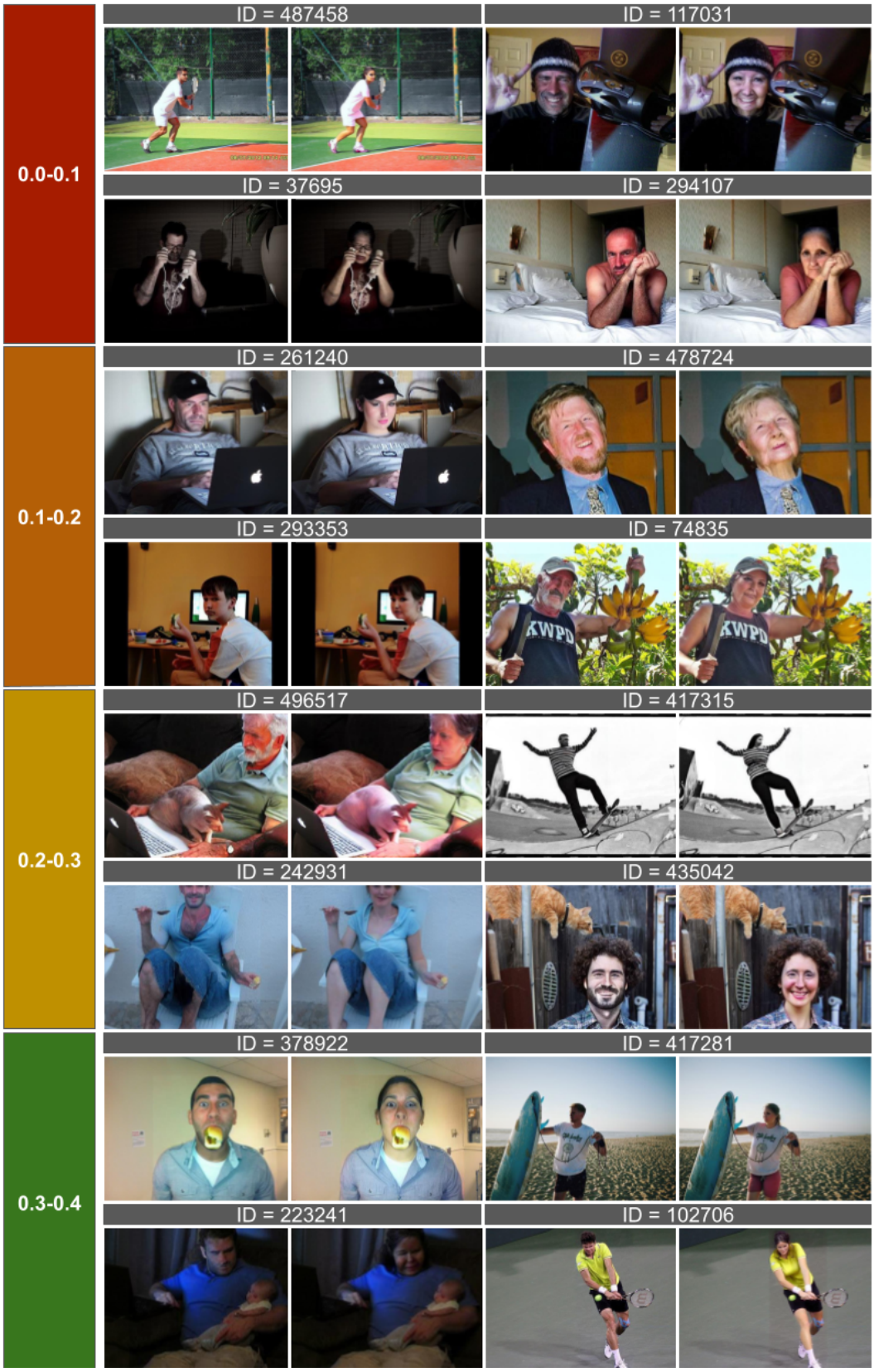
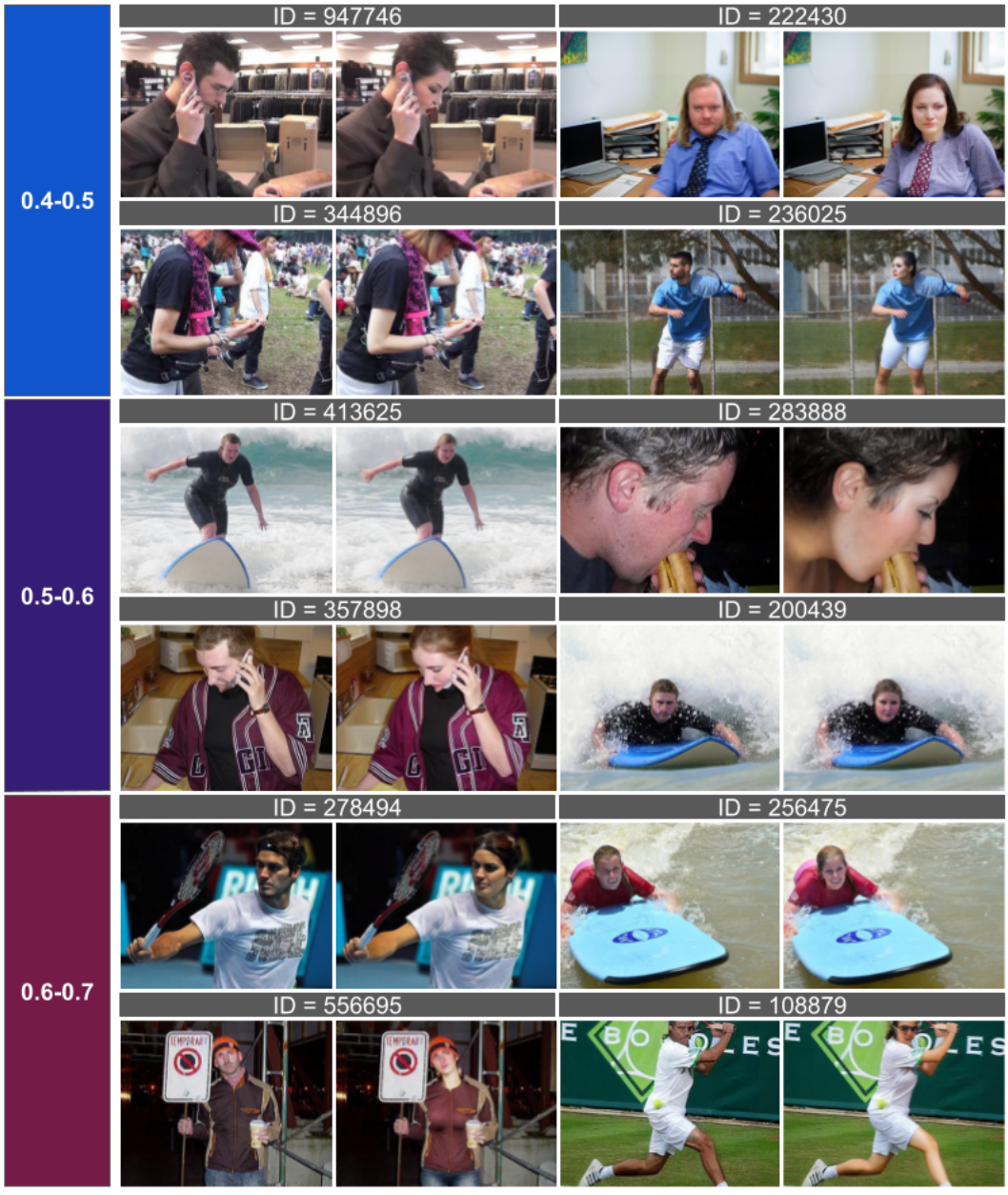


Figure 1: Randomly selected examples of GENSYNTH images showing a comparison to the original COCO image and the weak baseline GENSWAP.

Figure 2: 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.

REFERENCES

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.

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

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.

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.

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. URL `https://doi.org/10.5281/zenodo.5143773`.

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.

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

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.

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.

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. ISSN 0001-0782.

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.

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.