

Exploring Correlated Facial Attributes in Text-to-Image Models: Unintended Consequences in Synthetic Face Generation

Sander De Coninck, Sam Leroux, Pieter Simoens

IDLab, Department of Information Technology at Ghent University - imec

Technologiepark 126, B-9052 Ghent, Belgium

{sander.deconinck, sam.leroux, pieter.simoens}@ugent.be

Abstract

Text-based face manipulation is a powerful tool for generating and editing facial attributes in synthetic data, particularly for addressing imbalances in biometric datasets and helping to build fair and privacy-aware systems. However, controlling specific attributes remains challenging, as modifying one attribute often leads to unintended changes in others, potentially introducing new biases. In this paper, we investigate the cause of these unintended changes, hypothesizing that they stem from attribute correlations in the training data. By analyzing three face attribute datasets (CelebA, MAAD-Face, FFText-HQ), we identify common patterns of attribute co-occurrence and assess the behavior of three state-of-the-art manipulation models (ManiCLIP, FFCLIP, and DeltaEdit) using attribute classifiers to detect unwanted modifications. We found two types of errors: one where more attributes were altered than intended, and another where changes failed to occur due to dependencies on the original attributes of the face. While we hypothesized that these manipulation issues might correlate with dataset attribute correlations, our analysis using the Matthews Correlation Coefficient found no statistically significant relationships, suggesting that other factors may contribute to these unintended changes. This work can serve as a starting point for further research into the underlying causes of attribute entanglement in face manipulation models, helping to improve understanding and address related challenges.

1. Introduction

Text-based face generation has garnered increasing attention in recent years [18], as it enables the manipulation of biometric, demographic, and non-permanent facial attributes such as hairstyle, earrings, or lipstick. These models are now being utilized in the generation of synthetic data [1], offering an efficient means to create new data that adheres to a desired distribution. Many face-related

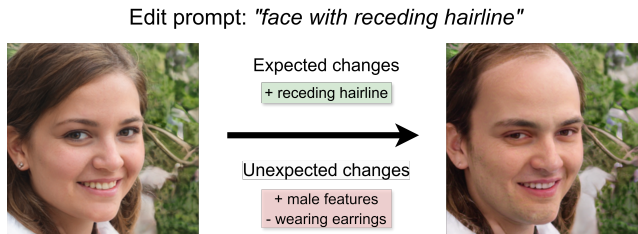


Figure 1. Example of unintended changes when editing face features through prompting text-to-image models. If applied for data augmentation, these effects could introduce new biases into datasets.

applications suffer from dataset imbalances, with certain demographic groups being underrepresented or even absent [32]. Given the high cost and impracticality of manual data collection, face manipulation models provide a cost-effective solution for augmenting datasets. This capability is particularly promising in the development of fair applications [26, 27, 37], enhancing privacy protection [5, 20, 21], and supporting data augmentation [6, 19, 39] within biometric systems. However, controlling individual attributes in synthetic data remains a significant challenge, as modifying one attribute often causes unintended changes in others [7, 25]. For instance, attempting to change a hair-related property of a synthetic face may unintentionally alter the perceived gender or age, an example can be seen in Figure 1. As these models are increasingly used for data augmentation, the potential for them to introduce new biases is a concerning issue. If left unchecked, this could reinforce or exacerbate existing imbalances rather than mitigate them.

In this paper, we explore how and why these unintended attribute changes occur. We hypothesize that these changes stem from correlations between facial attributes present in the training data. In real-world face datasets, attributes such as age, gender, and facial hair are often interrelated, reflecting natural patterns of co-occurrence. Generative models trained on such data may inadvertently encode these cor-

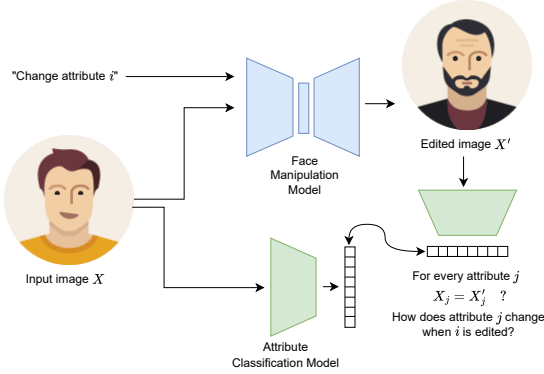


Figure 2. Workflow of our main experiment. We identify unintended changes in face manipulation models by applying attribute classifiers to both the original and edited images and verifying which attributes have been changed.

relations, leading to entangled attribute representations that make precise manipulation challenging.

Our investigation focuses on the relationship between attribute correlations in face datasets and the resulting behavior of text-to-image models during attribute manipulation. We begin by conducting a comprehensive analysis of existing correlations within three prominent face attribute datasets: CelebA [14], MAAD-Face [31], and FFText-HQ [30]. This analysis aims to uncover patterns of highly correlated attributes of faces, as well as any anomalies or unexpected correlations. Furthermore, we examine whether these correlations are consistent across the different datasets or if significant disparities emerge.

In the next step, we conduct a detailed evaluation of state-of-the-art face manipulation models, focusing on the occurrence of unintended attribute modifications during the editing process. To detect these unwanted changes, we employ attribute classification models, as outlined in Figure 2. We further investigate whether different models share common manipulation flaws, providing a comparative analysis of their behaviors. Additionally, we explore how pre-existing attributes influence the success of a given manipulation. For example, we assess whether adding earrings to a synthetic face is more effective for female faces compared to male ones. Lastly, we examine whether these unintended changes are related to the attribute correlations present in the training datasets, seeking deeper insights into the patterns that may drive these issues.

The paper is structured as follows: Section 2 covers related work on face manipulation and correlation studies. Section 3 details the experiment setup and tools. We present the main correlation and manipulation results in Section 4 and 5. Finally, we conclude in Section 6 and discuss future work in Section 7.

2. Related works

2.1. Face Manipulation and Correlations

Correlation is a widely studied problem in the domain of fair ML research, mainly focusing on spurious correlations in the dataset through societal biases or dataset construction. For example, for the CelebA dataset, it is known that the *Attractive* attribute is biased towards women and that a majority of the women have blonde hair. Rajabi et al. [25] demonstrate that mitigating bias between gender and attractiveness for the CelebA dataset influences classification of several other attributes related to attractiveness as well. Similarly, Denton et al. [7] found that manipulating non-hair attributes such as heavy makeup resulted in a change in hair style and/or color. Zhang et al. [38] studied attribute correlations and their effect upon adversarial learning and found them to be highly indicative of the privacy-utility tradeoff. Few methods have been proposed to combat attribute entanglement [29, 36] but none of these investigates the underlying causes of entanglement.

2.2. Manipulation of Face Attributes using Text-to-Image Models

StyleGAN’s [12] editable latent space, when combined with CLIP [24], enables precise, text-driven control over image features. Recent advances in GAN-based models have greatly enhanced facial image manipulation. ManiCLIP [35] introduces a method for multi-attribute facial editing via text prompts, decoupling attribute control to avoid irrelevant changes and eliminating the need for test-time optimization. AnyFace++ [30] supports flexible, open-world text-to-face generation, leveraging CLIP and a memory module for precise and diverse edits based on complex descriptions. DeltaEdit [15] uses a text-free training method, aligning CLIP’s visual and textual embedding differences to guide facial manipulation. It predicts StyleGAN’s editing directions from text variations, enabling efficient zero-shot inference without additional tuning. FFCLIP [40] offers a versatile approach for free-form text prompts by mapping CLIP embeddings to StyleGAN’s latent space, allowing one model to handle multiple semantic meanings in image edits. Finally, StyleEntity [8] uses named entities during training to allow zero-shot image manipulation with unseen prompts. Its Prompt Ensemble Latent Averaging (PELA) technique averages manipulation directions, reducing noise and ensuring stable edits.

3. Experiment setup

3.1. Data

We conduct our experiments using three publicly available facial image datasets (see Table 1 for an overview): CelebA [14], FFText-HQ [30] (a labeled subset of

FFHQ [11]), and MAAD-Face [31]. The CelebA dataset contains 202k images of 10k celebrities, labeled with 40 binary attributes such as Young, Eyeglasses, and Smiling. The FFHQ dataset itself does not include attribute labels, but Sun et al. [30] introduced FFText-HQ, a subset of 20k FFHQ images annotated with binary and multi-class attributes covering specific facial regions like Hair, Eyes, and Mouth, as well as broader attributes such as Age and Gender. Lastly, the MAAD-Face dataset provides labeled annotations for 47 binary attributes based on a subset of the VGGFace2 dataset [2], consisting of 3.3M images of 9k individuals. We excluded the LFW [9] dataset as its attributes have been proven to have a low correspondence with human annotators [31].

Table 1. Description of used face datasets, the number of attributes, and their types.

Dataset	Number of Attributes	Attribute Type
CelebA [14]	40	Binary
FFText-HQ [30]	26	Binary & Multiclass
MAAD-Face [31]	47	Binary

3.2. Categorical Correlation Metrics

To assess the correlation between categorical variables, we use three metrics: the **Matthews Correlation Coefficient (MCC)**, **Cramér’s V**, and the **Uncertainty Coefficient (Theil’s U)**.

Matthews Correlation Coefficient (MCC) [17] is the primary metric for binary classification, offering a balanced measure of association. It is particularly useful for imbalanced datasets, where accuracy may be misleading. The MCC is defined as:

$$\text{MCC} = \frac{TP \cdot TN - FP \cdot FN}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (1)$$

where TP , TN , FP , and FN are true positives, true negatives, false positives, and false negatives, respectively. The MCC ranges from -1 to 1 , with 1 indicating perfect prediction, 0 random prediction, and -1 total disagreement.

For non-binary categorical variables, we use **Cramér’s V** [3], a symmetric measure based on Pearson’s chi-squared statistic. It is calculated as:

$$V(X, Y) = \sqrt{\frac{\frac{\chi^2}{n}}{\min(k-1, r-1)}} \quad (2)$$

where χ^2 is Pearson’s chi-squared statistic, n is the number of samples, and k and r are the number of categories of X and Y . A value of $V = 0$ indicates no association, and values close to 1 indicate strong association.

The Uncertainty Coefficient (Theil’s U) [33] measures the reduction in uncertainty about one variable given knowledge of another. It is asymmetric and ranges from 0 to 1 . The formula is:

$$U(X|Y) = \frac{H(X) - H(X|Y)}{H(X)} = \frac{I(X; Y)}{H(X)} \quad (3)$$

where H represents entropy, and $I(X; Y)$ denotes mutual information. $U(X|Y)$ represents the percentage reduction in uncertainty about X given Y , providing an intuitive measure of informational dependency between categorical variables.

3.3. Face Attribute Classification

To evaluate the impact of data manipulation techniques, we require accurate models for classifying attribute values. For the CelebA dataset, several pre-trained models are available, including SwinFace [22], Faceptor [23], and DMM-CNN [16]. However, no public classifiers exist for the MAAD-Face and FFText-HQ datasets, so we focus on CelebA attributes. In this paper, we use the SwinFace model due to its open-source implementation and available model weights. Additionally, we employ a standard ResNet50 model, trained as outlined by [11, 13].

To minimize the influence of classifier selection on our results, we employed both classifiers to measure attribute changes for each experiment and averaged the outcomes across both models.

3.4. Face Image Manipulation

We use the following image manipulation models: ManiCLIP [35], FFCLIP [40], and DeltaEdit [15]. We chose these models because they are state-of-the-art models, that provide both code and trained models. For each model, we perform text-based manipulations to edit one attribute of the image and use the attribute classifiers before and after the changes on the images. To achieve optimal manipulation results with these models, we use the same text prompts that were employed during their training for each model. We apply all edits as defined in the implementations, including those that don’t directly correspond to CelebA attributes, to ensure a comprehensive evaluation of the editing results.

ManiCLIP: This method works from self-initialized latents and is thus not compatible with real images. To test out the changes, we first randomly generate images using the latent codes provided by the authors and then perform the manipulation on these images. We try out 22 edits, of which 17 map directly to CelebA attributes.

FFCLIP: This method is applied to the inverse latents of the FFHQ dataset, which we acquired using encoder4editing (e4e) [34]. We try out 38 edits, of which 7 map directly to CelebA attributes.

DeltaEdit: Similar to FFCLIP, this method is applied to the inverse latents of FFHQ, acquired through e4e. We try out 16 edits, of which 7 map directly to CelebA attributes.

Both FFCLIP and DeltaEdit were trained using the FFHQ dataset. In contrast, ManiCLIP was not trained directly on a dataset; however, it utilized a StyleGAN network pretrained on FFHQ in conjunction with text prompts similar to the CelebA attributes.

4. Correlation study

In this section, we analyze face attribute datasets, focusing on the relationships between attributes. We first identify the strongest correlations within each dataset, then compare these correlations across datasets to determine whether they are common across datasets or specific to each one.

4.1. Analysis of Strong Correlations in Face Attribute Datasets

Figure 3 displays the 20 strongest correlations between attributes across all datasets and metrics, including a classification of strength. It is important to note that the Uncertainty Coefficient is asymmetric, meaning some attribute pairs may appear twice in the list. Many of the strongest correlations are logical and common across different datasets. For example, strong correlations frequently occur between attributes like Male¹ and facial hair attributes like Goatee, Mustache and No Beard, or strong correlations between Wearing Lipstick and Heavy Makeup or wrinkles and age.

Some of the strong correlations we identified could be problematic if they are carried over into face manipulation tasks. For instance, attributes such as Male, Wearing Lipstick, Wearing Earrings, Hair Length, and Hair Color show correlations that may reinforce gender stereotypes. When making a face appear more female, the model might automatically add lipstick, reinforcing biased associations. Similarly, the correlations between Wearing Lipstick and Wearing Earrings, or between Hair Color and Hair Length, reflect dependencies that, in theory, should be independent of one another.

While many of these correlations appear consistently in each dataset, we will verify their similarity in the next section.

4.2. Consistency of Correlation Strengths Across Datasets

To assess whether the correlations observed in individual datasets are consistent across all datasets, we converted

¹In the CelebA and MAAD-Face datasets, perceived gender is represented by the attribute Male, where a value of 1 indicates individuals categorized as men, and a value of 0 refers to those who are not, which we refer to as women for simplicity.

Table 2. This table highlights the most significant variations in Matthew’s Correlation Coefficient observed between the datasets.

Combination	Matthew’s Correlation Coefficient		
	CelebA	MAAD	FFText-HQ
5 O’Clock Shadow, No Beard	-0.53	-0.05	-0.66
Mustache, No Beard	-0.45	0.32	-0.27
Heavy Makeup, Male	-0.67	-0.78	-0.21
Heavy Makeup, Wavy Hair	0.32	0.68	0.04
Wavy Hair, Wearing Earrings	0.12	0.62	0.09
Goatee, Sideburns	0.51	0.28	-0.03
Goatee, No Beard	-0.57	0.35	-0.23
Smiling, Mouth Slightly Open	0.54	0.19	-0.31
Mustache, Goatee	0.45	0.91	-0.02
Smiling, High Cheekbones	0.68	0.62	-0.04

them into a unified format. We standardized the datasets to align with the format of CelebA, the most widely used dataset, and focused on attributes that were present in all three datasets. In total, 30 attributes (out of CelebA’s original 40) were shared across the datasets.

To examine if similar relationships exist, we applied Pearson and Kendall rank correlations to compare the inter-attribute correlations in each dataset. The findings, shown in Figure 4, indicate that the strongest correlations occur between the CelebA and MAAD-Face datasets, while the correlations with FFText-HQ, though still significant, are slightly weaker. One possible explanation for this difference is the nature of the underlying data: CelebA and MAAD-Face focus on public figures and celebrities, whereas FFText-HQ draws from Flickr, potentially offering a more general population sample.

Some of the largest differences in correlation strengths are highlighted in Table 2. A significant portion of these differences relates to facial hair attributes, likely due to variations in dataset labelling guidelines. In CelebA, multiple facial hair attributes can co-occur, whereas in the other datasets, many are mutually exclusive. Other notable differences include a weaker correlation between Heavy Makeup and both Male and Wavy Hair. Additionally, correlations in the FFText-HQ dataset are generally smaller, which may again be attributed in the domain difference.

5. Analysis of Face Attribute Manipulation and Its Relationships with Dataset Correlations

Here, we present the results of the face attribute manipulation experiments. We analyze both the unintended and intended changes produced by applying face editing using the three different models. Specifically, we investigate whether any unintended changes are consistent across the models and whether these changes align with the correlations identified in Section 4. Additionally, we explore instances where attribute changes are not fully applied, ex-

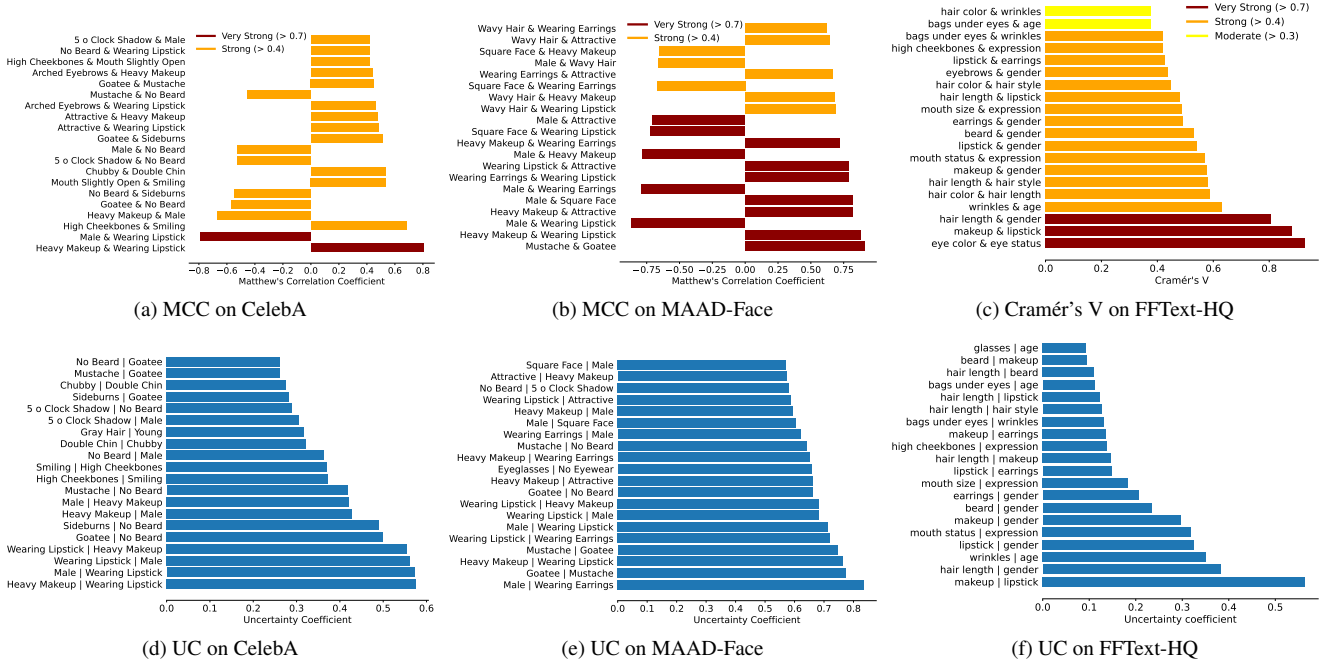


Figure 3. Top 20 strongest correlations calculated on CelebA, MAAD-Face and FFText-HQ datasets using the MCC/Cramer’s V and Uncertainty Coefficient (UC) correlation metrics. Several strong correlations are found across the datasets. Note that UC is an asymmetric metric, with the correlation for attribute x given y expressed as $x|y$

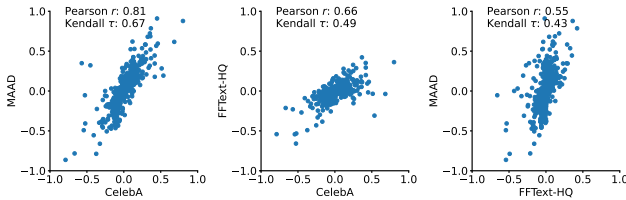


Figure 4. **Comparison of Inter-Dataset Correlations.** Each point represents the MCC of an attribute pair across two datasets. Additionally, we present the Pearson and Kendall rank correlations for all points. All correlations were found to be significant with $p < 0.001$. The correlations between datasets are apparent, though they are notably weaker for the FFText-HQ dataset.

amining whether these outcomes are related to the original appearance of the faces.

5.1. Unintended Effects of Face Manipulation

The results of our main experiment are presented in Figure 5 for each manipulation model, with a larger version available in the supplementary material. Below, we highlight some key findings for each model.

ManiCLIP: This model shows a significant number of unintended changes, most notably in the `Brown Hair` category. We observed that many output images displayed brown hair, even when the manipulation was unrelated to hair color. Other notable patterns include: applying the

Young transformation frequently resulted in subjects with blond hair; adding a receding hairline increased masculine features; and the `Big Lips` manipulation often introduced lipstick and heavy makeup. Examples of these edits and their effects can be seen in Figure 6a.

FFCLIP: Despite being the oldest model evaluated, FFCLIP supports the widest range of edits, though some manipulations showed little or no effect. For example, altering a subject to appear male at age 40 or 80, closing the eyes, or adding glasses often had minimal impact. Unintended effects were also present, such as the `cornrows hairstyle` edit often leading to blond hair, or adding a beard making the subject look significantly older.

DeltaEdit: DeltaEdit demonstrates the fewest unintended changes overall. However, we also observed that certain intended modifications had minimal impact on the desired attributes. For instance, adding a beard resulted in only slight adjustments to facial hair features, and the same applied to adding earrings. Our manual review of the editing outcomes suggests that these manipulations are dependent on the attributes of the original image. For example, attempts to add a beard were less effective when the subject appeared female. We will explore this further in the following section.

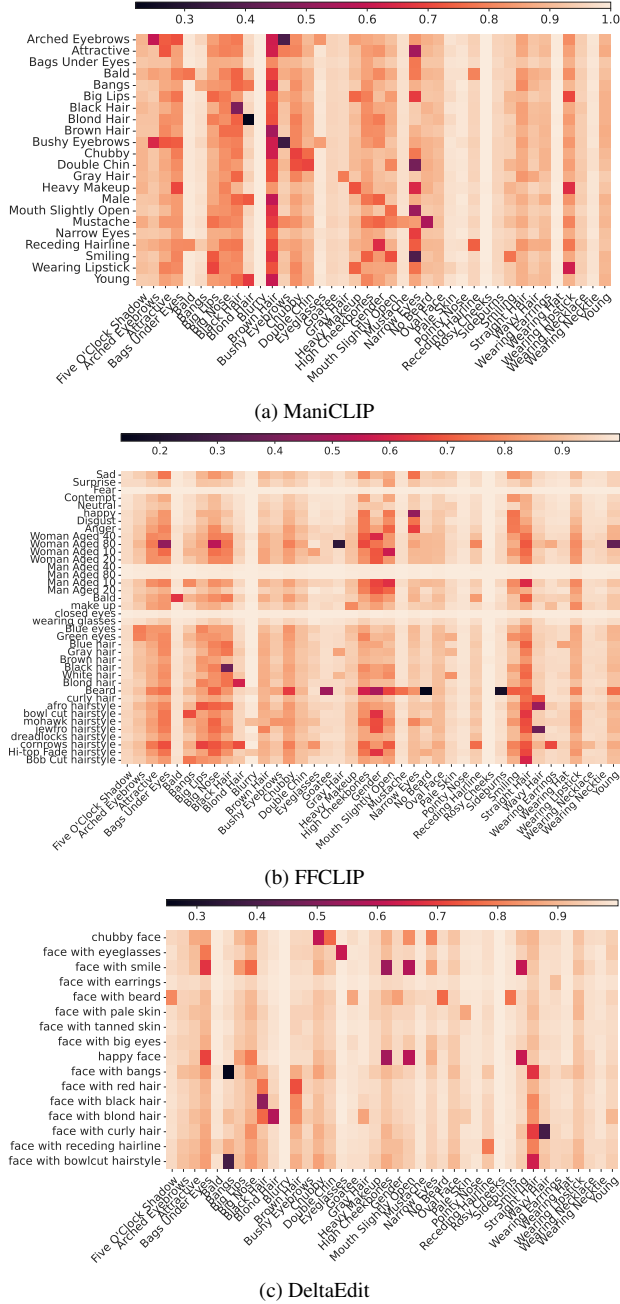


Figure 5. **Attribute Manipulation Results.** Each row presents face edits targeting specific attributes, while the columns reflect the degree of change in each attribute after manipulation (with a value of 1 indicating no change). Ideally, only the targeted attribute should exhibit a noticeable change, indicated by a value significantly lower than 1, while all other attributes should remain unchanged at 1.

5.2. Influence of Original Attributes on Manipulation Success

In DeltaEdit, we observed that some attribute modifications were significantly less successful than others, which we hypothesize is due to characteristics inherent in the original image. To investigate this further, we analyzed how each attribute, referred to as a confounder, affected the frequency and success of the modifications. Some notable examples of significant discrepancies in effectiveness are shown in Figure 7a, along with additional examples in Figure 6b. For instance, it appears more challenging to add eyeglasses to a young person or earrings to a man. Interestingly, adding earrings is rarely successful, even when applied to women.

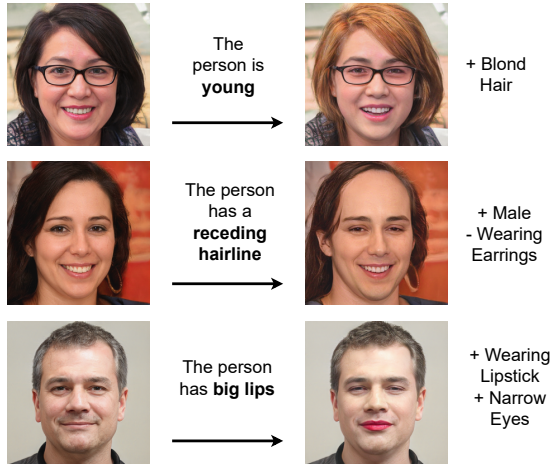
We observed similar trends in ManiCLIP, as illustrated in Figure 7b. For example, giving a man a larger nose or making him bald seems easier compared to performing the same alterations on a woman. Additionally, individuals already wearing eyeglasses are more likely to have their hair turned gray. However, these patterns were less pronounced in FFCLIP, likely due to the less direct relationship between the suggested edits and the classified attributes in this model.

5.3. Exploring the Relationship Between Attribute Manipulation and Dataset Correlations

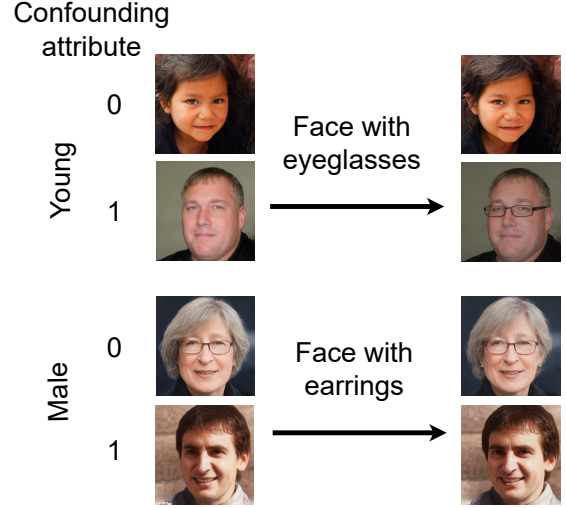
Finally, we attempt to relate our manipulation results to the correlations identified in Section 4. We focus on edits that can be directly mapped to attributes in the common set of our datasets. Specifically, we aim to correlate the changes in attribute Y resulting from the manipulation of attribute X (as shown in Figure 5) with the correlations between X and Y found in our datasets. In theory, we would expect a negative correlation, as higher correlations between X and Y would suggest greater changes when X is manipulated.

Furthermore, since most models were trained on the FFHQ dataset, either directly or via a pretrained module, we anticipated the strongest relationship would be with the FFText-HQ dataset, as it is a subset of FFHQ. However, no statistically significant results were observed for the MCC or UC values across any of the face attribute datasets. Scatterplots illustrating the relationship between manipulation and correlation for the MCC metric are presented in Figure 8.

A possible explanation for this fact is that looking at attribute correlation at the label level may be overly abstract. Instead, it might be more meaningful to do the analysis in latent or pixel space, for example by using interpretability techniques such as XRAI [10], GradCAM [28] or a region-based attribution technique [4].



(a) A face edit intended to target only one attribute leads to unexpected changes. Examples shown are generated with ManiCLIP.



(b) Impact of Original Attributes on Manipulation. Changes are less likely to occur if a certain attribute is already present in the original image. Examples generated with DeltaEdit.

Figure 6. Examples of two possible issues when using face image manipulation methods

6. Conclusion

In this study, we investigated the relationship between face attribute correlations across multiple datasets and how these correlations impact the effectiveness of face manipulation models. By analyzing the strongest correlations found in the datasets and comparing them across CelebA, MAAD-Face, and FFText-HQ, we observed that while many correlations were logical and consistent, others could propagate biases when embedded in facial manipulation models, such as the correlation between *Wearing Earrings* and *Male*.

In our experiments with three face manipulation models (ManiCLIP, FFCLIP, and DeltaEdit) we observed that certain attribute changes often led to unintended modifications. For instance, editing one feature, like hairstyle or age, sometimes altered unrelated attributes such as makeup or hair color. We also found that the effectiveness of manipulations sometimes depended on the original attributes of the face. Some edits, like adding eyeglasses or earrings, were less successful when certain confounding attributes, like gender or age, were present. These findings suggest that attribute correlations in the models contribute to these unintended changes.

However, we found no statistically significant relationships when attempting to correlate the attribute correlations from the datasets with the manipulation results. Although we expected higher correlations to lead to more difficulties in independent manipulation, no clear pattern emerged, suggesting that other factors may play a more significant role in determining manipulation success.

Overall, our findings highlight the complexity of face attribute manipulation and the need for further research to understand how correlations in training data might influence manipulation outcomes. Future work could explore ways to reduce unintended biases and improve the predictability and accuracy of face editing models across diverse datasets and demographic groups.

7. Implications for Further Research & Future Work

The results of our experiment highlight two key findings: first, we observed cases where face manipulation led to unintended consequences, and second, we could not relate these outcomes to the correlations calculated using the MCC and Uncertainty coefficient. We hypothesize that these unintended effects may have broader implications for future research. Specifically, a “no changes” bias could inadvertently reveal information about the original subject, while excessive changes might negatively impact the generalization ability of manipulation models, making them less robust to less common cases (e.g., men wearing earrings).

Since we could not link the observed effects to the correlations derived from the MCC and Uncertainty coefficient, we propose an alternative approach: calculating correlations by directly examining the data and utilizing interpretability techniques. Methods such as those explored in [4, 10, 28] could provide deeper insights into the interconnectedness of attributes by looking at the regions in images they pertain to, enabling a more nuanced understanding of how manipulation impacts different face features.

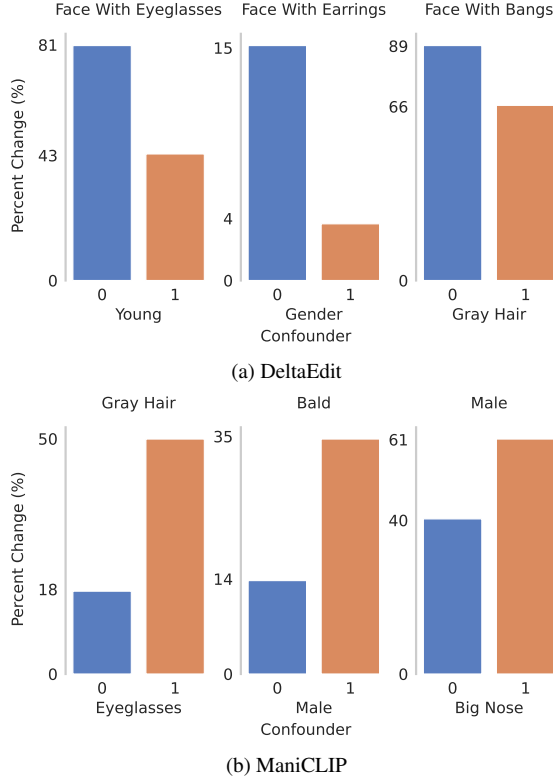


Figure 7. **Examples of Manipulation Effectiveness Based on Attribute Presence (Confounder).** This shows how the effectiveness of manipulations varies depending on whether a certain attribute (confounder) is present in the original image.

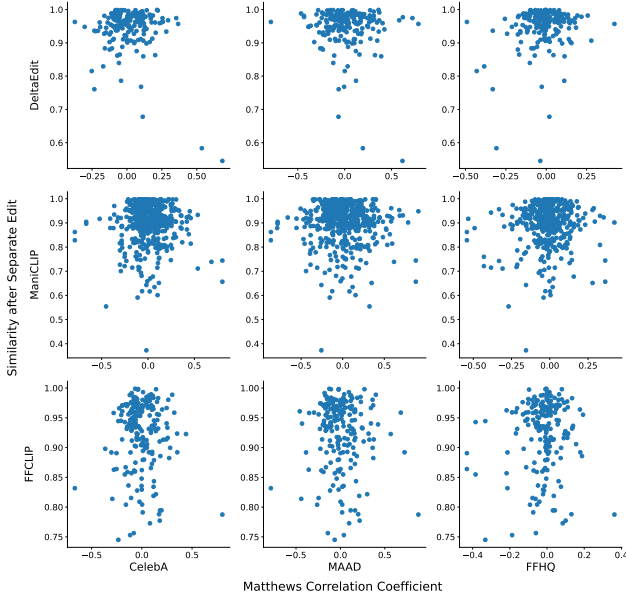


Figure 8. **Relationship Between Dataset Correlations (MCC) and Manipulation Results.** We observed no statistically significant correlations between the correlations present in the dataset and the manipulation outcomes.

We view this paper as an initial step in exploring this area, and further research is needed to address several key questions. Future work should explore different metrics for calculating correlations to determine if they can explain the scenarios in which certain issues arise. Additionally, investigating the effects of manipulating multiple attributes simultaneously, as opposed to focusing on individual attribute changes, would be an intriguing direction for further study.

Acknowledgments

Sander De Coninck receives funding from the Special Research Fund of Ghent University under grant no. BOF22/DOC/093. This work received financial support from the Flanders AI Research Program (FAIR).

References

- [1] Georgia Baltso, Ioannis Sarridis, Christos Koutlis, and Symeon Papadopoulos. Sdfd: Building a versatile synthetic face image dataset with diverse attributes. *arXiv preprint arXiv:2404.17255*, 2024. 1
- [2] Qiong Cao, Li Shen, Weidi Xie, Omkar M. Parkhi, and Andrew Zisserman. VGGFace2: A dataset for recognising faces across pose and age. In *International Conference on Automatic Face and Gesture Recognition*, 2018. 3
- [3] Harald Cramér. Mathematical methods of statistics (PMS-9), volume 9. In *Mathematical Methods of Statistics (PMS-9), Volume 9*. Princeton university press, 2016. 3
- [4] Sander De Coninck, Sam Leroux, and Pieter Simoens. Mitigating bias using model-agnostic data attribution. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 235–243, 2024. 6, 7
- [5] Sander De Coninck, Wei-Cheng Wang, Sam Leroux, and Pieter Simoens. Selective manipulation of disentangled representations for privacy-aware facial image processing. *arXiv preprint arXiv:2208.12632*, 2022. 1
- [6] Ivan DeAndres-Tame, Ruben Tolosana, Pietro Melzi, Ruben Vera-Rodriguez, Minchul Kim, Christian Rathgeb, Xiaoming Liu, Luis F Gomez, Aythami Morales, Julian Fierrez, et al. Second frcsyn-ongoing: Winning solutions and post-challenge analysis to improve face recognition with synthetic data. *arXiv preprint arXiv:2412.01383*, 2024. 1
- [7] 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. 1, 2
- [8] Zhida Feng, Li Chen, Jing Tian, JiaXiang Liu, and Shikun Feng. Named entity driven zero-shot image manipulation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9110–9119, 2024. 2
- [9] Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, October 2007. 3

- [10] Andrei Kapishnikov, Tolga Bolukbasi, Fernanda Viégas, and Michael Terry. Xrai: Better attributions through regions. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4948–4957, 2019. 6, 7
- [11] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4401–4410, 2019. 3
- [12] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and improving the image quality of stylegan. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8110–8119, 2020. 2
- [13] Ji Lin, Richard Zhang, Frieder Ganz, Song Han, and Jun-Yan Zhu. Anycost gans for interactive image synthesis and editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14986–14996, 2021. 3
- [14] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015. 2, 3
- [15] Yueming Lyu, Tianwei Lin, Fu Li, Dongliang He, Jing Dong, and Tieniu Tan. Deltaedit: Exploring text-free training for text-driven image manipulation, 2023. 2, 3
- [16] Longbiao Mao, Yan Yan, Jing-Hao Xue, and Hanzi Wang. Deep multi-task multi-label cnn for effective facial attribute classification. *IEEE Transactions on Affective Computing*, 13(2):818–828, 2020. 3
- [17] Brian W Matthews. Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochimica et Biophysica Acta (BBA)-Protein Structure*, 405(2):442–451, 1975. 3
- [18] Andrew Melnik, Maksim Miasayedzenkau, Dzianis Makaravets, Dzianis Pirshtuk, Eren Akbulut, Dennis Holzmann, Tarek Renusch, Gustav Reichert, and Helge Ritter. Face generation and editing with stylegan: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 1
- [19] Pietro Melzi, Christian Rathgeb, Ruben Tolosana, Ruben Vera-Rodriguez, Dominik Lawatsch, Florian Domin, and Maxim Schaubert. Gandiffface: Controllable generation of synthetic datasets for face recognition with realistic variations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops*, pages 3086–3095, October 2023. 1
- [20] Vahid Mirjalili, Sebastian Raschka, and Arun Ross. Flowsan: Privacy-enhancing semi-adversarial networks to confound arbitrary face-based gender classifiers. *IEEE Access*, 7:99735–99745, 2019. 1
- [21] Vahid Mirjalili, Sebastian Raschka, and Arun Ross. Privacynet: Semi-adversarial networks for multi-attribute face privacy. *IEEE Transactions on Image Processing*, 29:9400–9412, 2020. 1
- [22] Lixiong Qin, Mei Wang, Chao Deng, Ke Wang, Xi Chen, Jiani Hu, and Weihong Deng. Swinface: a multi-task transformer for face recognition, expression recognition, age estimation and attribute estimation. *IEEE Transactions on Circuits and Systems for Video Technology*, 2023. 3
- [23] Lixiong Qin, Mei Wang, Xuannan Liu, Yuhang Zhang, Wei Deng, Xiaoshuai Song, Weiran Xu, and Weihong Deng. Faceptor: A generalist model for face perception. *arXiv preprint arXiv:2403.09500*, 2024. 3
- [24] 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. 2
- [25] Amirarsalan Rajabi, Mehdi Yazdani-Jahromi, Ozlem Ozmen Garibay, and Gita Sukthankar. Through a fair looking-glass: mitigating bias in image datasets. In *International Conference on Human-Computer Interaction*, pages 446–459. Springer, 2023. 1, 2
- [26] 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. 1
- [27] 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. 1
- [28] Ramprasaath R Selvaraju, Michael Cogswell, Abhishek Das, Ramakrishna Vedantam, Devi Parikh, and Dhruv Batra. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, pages 618–626, 2017. 6, 7
- [29] Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. Interpreting the latent space of gans for semantic face editing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9243–9252, 2020. 2
- [30] Jianxin Sun, Qiyao Deng, Qi Li, Muyi Sun, Yunfan Liu, and Zhenan Sun. Anyface++: A unified framework for free-style text-to-face synthesis and manipulation. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 2, 3
- [31] Philipp Terhöst, Daniel Fährmann, Jan Niklas Kolf, Naser Damer, Florian Kirchbuchner, and Arjan Kuijper. Maad-face: A massively annotated attribute dataset for face images. *IEEE Trans. Inf. Forensics Secur.*, 16:3942–3957, 2021. 2, 3
- [32] Philipp Terhöst, Jan Niklas Kolf, Marco Huber, Florian Kirchbuchner, Naser Damer, Aythami Morales Moreno, Julian Fierrez, and Arjan Kuijper. A comprehensive study on face recognition biases beyond demographics. *IEEE Transactions on Technology and Society*, 3(1):16–30, 2022. 1
- [33] Henri Theil. On the Estimation of Relationships Involving Qualitative Variables. *American Journal of Sociology*, 76(1):103–154, July 1970. 3
- [34] Omer Tov, Yuval Alaluf, Yotam Nitzan, Or Patashnik, and Daniel Cohen-Or. Designing an encoder for stylegan image manipulation. *arXiv preprint arXiv:2102.02766*, 2021. 3
- [35] Hao Wang, Guosheng Lin, Ana García del Molino, Anran Wang, Jiashi Feng, and Zhiqi Shen. Maniclip: Multi-attribute face manipulation from text. *International Journal of Computer Vision*, pages 1–17, 2024. 2, 3

- [36] Ran Yi, Teng Hu, Mengfei Xia, Yizhe Tang, and Yong-Jin Liu. Feditnet++: Few-shot editing of latent semantics in gan spaces with correlated attribute disentanglement. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024. 2
- [37] Seyma Yucer, Samet Akçay, Noura Al-Moubayed, and Toby P Breckon. Exploring racial bias within face recognition via per-subject adversarially-enabled data augmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 18–19, 2020. 1
- [38] Likun Zhang, Yahong Chen, Ang Li, Binghui Wang, Yiran Chen, Fenghua Li, Jin Cao, and Ben Niu. Interpreting disparate privacy-utility tradeoff in adversarial learning via attribute correlation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 4701–4709, 2023. 2
- [39] Xinyue Zhu, Yifan Liu, Jiahong Li, Tao Wan, and Zengchang Qin. Emotion classification with data augmentation using generative adversarial networks. In *Advances in Knowledge Discovery and Data Mining: 22nd Pacific-Asia Conference, PAKDD 2018, Melbourne, VIC, Australia, June 3-6, 2018, Proceedings, Part III* 22, pages 349–360. Springer, 2018. 1
- [40] Yiming Zhu, Hongyu Liu, Yibing Song, Ziyang Yuan, Xintong Han, Chun Yuan, Qifeng Chen, and Jue Wang. One model to edit them all: Free-form text-driven image manipulation with semantic modulations. *Advances in Neural Information Processing Systems*, 35:25146–25159, 2022. 2, 3

Exploring Correlated Facial Attributes in Text-to-Image Models: Unintended Consequences in Synthetic Face Generation

Supplementary Material

A. Common attributes between face attribute datasets

Common attributes (30 total): Receding Hairline, Big Nose, Heavy Makeup, Pointy Nose, Smiling, Male, Bangs, Mustache, Brown Hair, Eyeglasses, Arched Eyebrows, Wearing Lipstick, Black Hair, Goatee, Blond Hair, Young, Bald, Mouth Slightly Open, Sideburns, 5 o Clock Shadow, Wavy Hair, Wearing Hat, High Cheekbones, Bags Under Eyes, No Beard, Gray Hair, Wearing Earrings, Bushy Eyebrows, Oval Face, Rosy Cheeks

B. Manipulation Results Full Figures

Full versions of the figures shown in Section 5 can be seen in Figure B.9 for DeltaEdit, Figure B.10 for ManiCLIP and Figure B.11 for FFCLIP

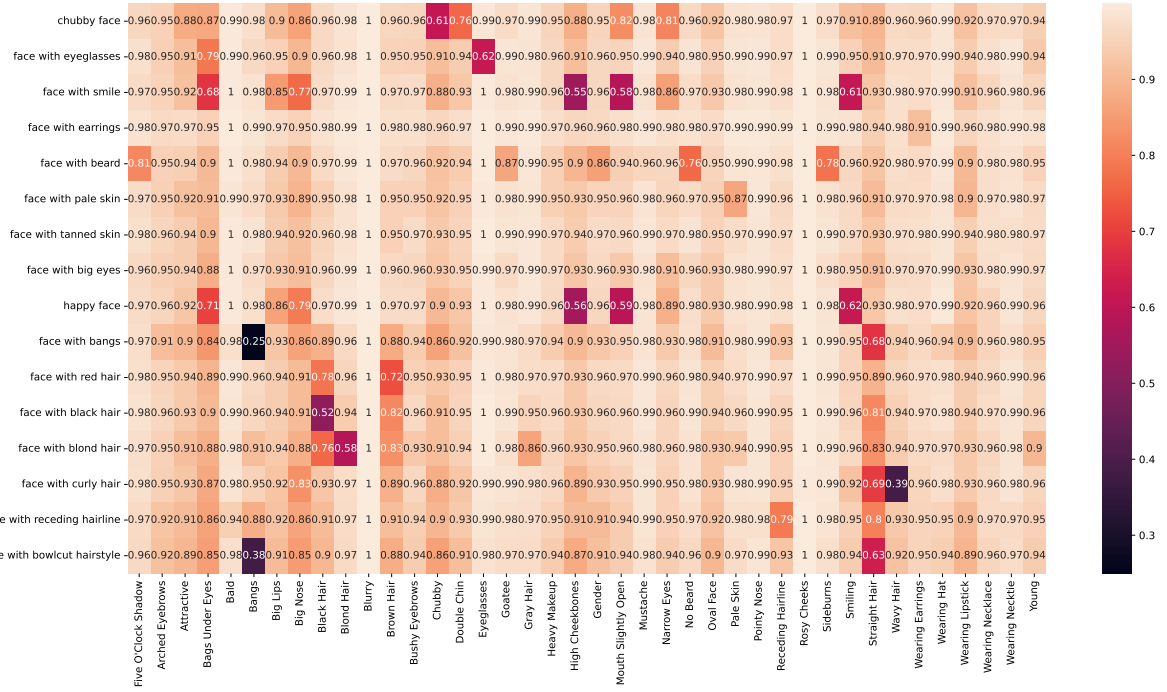


Figure B.9. Full manipulation results for DeltaEdit

