

A Large-scale Evaluation of Text-guided Models for Facial Editing

Rahul Nair
rnair21@asu.edu
Arizona State University
Tempe, AZ, USA

Saurav Pandit
saurav@prelight.ai
Prelight Inc
New York, USA

Hannah Kerner
hkerner@asu.edu
Arizona State University
Tempe, AZ, USA

Abstract

Facial appearance editing powers popular applications like FaceApp and Photoshop. Generative Adversarial Networks (GANs) and 3D Morphable Models (3DMMs) have been widely used for facial editing. GANs can perform varied facial edits (e.g., changing hair color, hairstyle), but often produce unstable edits. 3DMMs produce stable edits, but can only alter pose and facial expression. Recently, text-guided diffusion models like Nano Banana have become popular for image editing. Text-guided models are a compelling alternative to GANs and 3DMMs since they can produce both stable and varied image edits. While text-guided models have been widely tested for whole-scene edits (e.g., “make the woman play a guitar”), they have not been comprehensively tested for facial editing. We conducted the first large-scale evaluation (~ 1M images evaluated) of six popular text-guided models on a sequential facial editing task. We present Face-Edit-Attributes, the largest collection of 169 facial editing attributes focused on hair, accessories, and pose edits. We compared model performance using two popular celebrity face datasets: CelebA and CelebSET. Our results show that most models performed hair and accessory edits well, but struggled with editing pose. All models over-edit (e.g., changing hair color when asked only to change the hairstyle). We also evaluated demographic biases in each model. Our results show surprising biases in overediting: almost all models created more overedits for dark-skinned male faces and old faces. The code and data for our results (including our repository of ~ 1M images) can be accessed [here](#).

CCS Concepts

• **Social and professional topics** → **Cultural characteristics**.

Keywords

AI Evaluation, Generative AI Models, Societal Bias, Demographic Bias, Image Editing, Facial Editing.

ACM Reference Format:

Rahul Nair, Saurav Pandit, and Hannah Kerner. 2026. A Large-scale Evaluation of Text-guided Models for Facial Editing. In *Proceedings of the 34th ACM International Conference on Multimedia (MM '26)*, November 10–14, 2026, Rio de Janeiro, Brazil. ACM, New York, NY, USA, 29 pages. <https://doi.org/10.1145/3767308.3836090>

1 Introduction

Facial appearance editing is an important computer vision task that powers popular applications such as FaceApp [2] and Photoshop

[3]. The goal is to perform controlled edits on a person’s face (e.g., changing a person’s head pose) without altering additional elements in the image (e.g., changing hair color, background objects) or modifying the person’s identity. Facial editing has widespread applications, such as virtual try-ons (e.g., adding glasses, hats, and new hair colors) and content creation.

In recent years, Generative Adversarial Networks (GANs) [22, 23, 42, 47, 74] and 3D-morphable models (3DMMs) [28, 31, 34, 40] have been widely used for facial appearance editing. Both models have their pros and cons. While GANs can perform varied edits (e.g., changing hair color, hairstyle), they often produce unstable edits. While 3DMMs produce stable edits, they can only edit head pose and facial expression. This is because 3DMMs rely on 3D facial renderings where only pose and expression can be controlled.

Recently, text-guided image editing diffusion models like Nano Banana [11], which allow users to edit images using prompts (e.g., “make the person look leftward”), have become popular for image editing tasks. Text-guided models provide a compelling alternative for facial editing to GANs and 3DMMs. Text-guided models are trained on a massive corpus of image-and-text pairs, enabling them to perform a much larger variety of facial edits compared to GANs (e.g., adding different facial hair styles, different hat styles). Text-guided models also produce consistent and photorealistic edits due to their use of diffusion models.

There are several key requirements when using text-guided models for real-world facial editing applications: (1) Models should be able to execute granular facial edits (e.g., adding specific facial hair styles like a handlebar mustache or hat types like a slouchy beanie hat). (2) Models should perform edits fairly on male and female faces (no gender bias), on light and dark-skinned faces (no skin-tone bias), and on young and old faces (no age bias). (3) Models should not overedit — for instance, if a model is prompted to edit a person’s hair color, it should not change their hairstyle.

Previous studies comprehensively evaluate text-guided models on their ability to execute generic whole-scene edits (e.g., making a person ride a dragon) [16, 32, 53, 63, 71, 73]. However, previous studies have not comprehensively evaluated text-guided models’ ability to perform granular facial editing tasks. They also have not examined bias in these models across demographic groups.

We present the first large-scale evaluation of six state-of-the-art text-guided image editing models (Flux-Dev, Nano Banana, Bagel-Edit, Flux-Kontext-Pro, Qwen-Edit, and SeedEdit 3.0) for the facial appearance editing task. To perform facial edits, we propose **Face-Edit-Attributes** (Figure 2), a systematic collection of 169 manually curated facial editing attributes that focus on three types of edits: Hair (changing hair color, hair style, and adding facial hair), Accessories (adding different types of hats, glasses), and Pose/Gaze (making the person look left, right). Face-Edit-Attributes is an order



This work is licensed under a Creative Commons Attribution 4.0 International License. *MM '26, Rio de Janeiro, Brazil*

© 2026 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-2213-4/2026/11

<https://doi.org/10.1145/3767308.3836090>

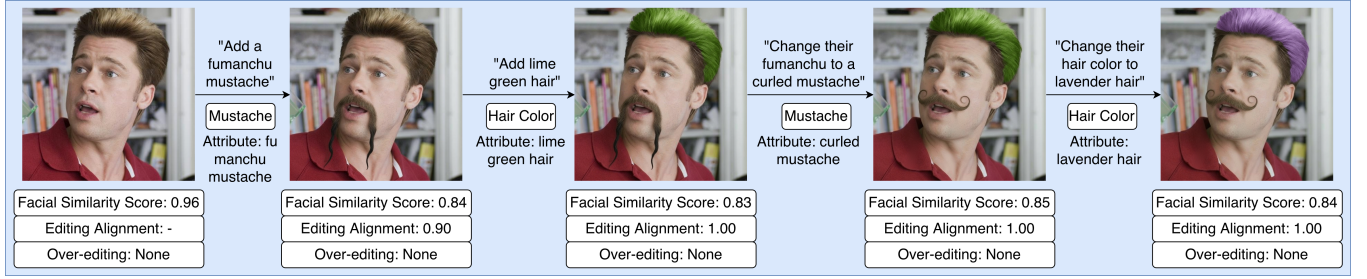


Figure 1: An example editing session from our study. Each image goes through a fixed sequence of 4 edits.

of magnitude larger than prior facial editing datasets (e.g., CelebA-dialog [35], CHATEDIT [24]).

We evaluated text-guided models on two popular celebrity face datasets: CelebA [49] and CelebSET [57]. To mimic practical facial editing use cases (e.g., retouching, virtual try-ons), we focused on multi-turn editing, where an image is edited using a sequence of editing instructions (Figure 1). For each image, we applied four editing instructions. We created separate evaluation sets for each edit type: Hair-Test, Accessories-Test, and Pose-Test. We also created a Multi-Axis-Test set, which is a mixture of these edit types. On each of these test sets, we compared overall model performance. We also evaluated whether models showed gender, skin-tone, and age bias.

We used 3 metrics for evaluation: (1) facial similarity score (FSS), which evaluates how well a person’s facial identity is preserved, (2) semantic consistency (SC), which evaluates whether the model executed the requested edit, and (3) we propose a new metric called the **overediting matrix**, which measures how many and which additional edits a model performed beyond the requested edit.

Our results show that most models correctly execute edits (high SC), while preserving > 70% of a person’s facial identity across the edit sequence (high FSS). However, performance is not the same across test sets — most models executed hair and accessory-based edits well, but they struggled with pose edits. All models overedit — for every 100 editing instructions (25 sequences), models created 25 overedits on average. The most common overedit patterns were associated with hair-edits — for instance, several models changed hair color when they were asked to edit the person’s hairstyle. Such overedit patterns are indicative of spurious correlations learned by models that future research should seek to mitigate.

In identity preservation (FSS) and in executing edits (SC), only Qwen-Edit and Flux-Dev showed consistent gender and skin-tone bias across test sets. We observed age bias in identity preservation — several models better preserved the facial identity of young faces. We also observed bias in overediting performance — all models created significantly more overedits when editing men (especially dark-skinned men), and when editing old people.

In summary, this paper makes the following contributions: (1) We conducted the first large-scale evaluation of text-guided editing models (around 1M images evaluated). (2) We propose Face-Edit-Attributes, the largest collection of 169 attributes to perform facial editing. (3) We systematically evaluated four types of facial edits: hair, accessories, pose, and multi-axis edits. (4) We propose a new metric, the Overediting Matrix, to count overedits. (5) We conduct the first study of demographic biases (gender, skin-tone, and age

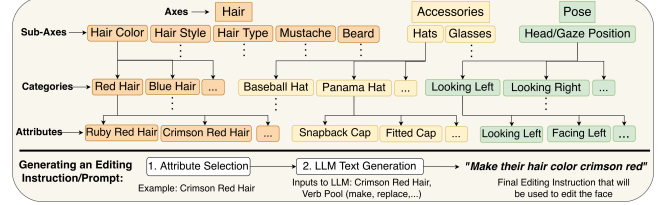


Figure 2: Face-Edit-Attributes: Leaf nodes (e.g., crimson red hair) are used to create the editing instructions.

bias) in facial editing, and show that models create more overedits when they edit male faces and old faces.

Summary of Related Work. : Current benchmarks for facial editing (e.g., MM-CelebA-HQ [66], CHAT-EDIT [24], CelebA-dialog [35]) have only 15 – 35 useful editing attributes.

Well-known image editing benchmarks (e.g., Magic Brush [71], Instruct-Pix-2-Pix [16], Omni-Edit [63]) have evaluated text-guided models on synthetic images (generated by models like DALL-E 3 [33] and Stable Diffusion [58]), on scene datasets like COCO [46] and RefCOCO [70], or on online image repositories like Pexels [4] and Flickr [69]. None of these benchmarks used face datasets (e.g., CelebA [49], CelebA-HQ [39]) for evaluation.

Many metrics have been proposed to measure overediting in models [14, 41, 45, 48, 55]. All of these metrics report overediting with a numerical score. This does not provide information on what overedits occurred in the image. We can modify DICE [14] (one of these metrics) to detect what overedits occurred. However, DICE cannot reliably identify fine-grained overedits (e.g., did the model overedit a hat or a Panama hat?).

For faces, most studies have examined demographic biases in tasks such as facial recognition, classification, and generation [18–20, 44]. To our knowledge, no work has extensively evaluated demographic biases in facial editing. For facial editing, only stereotype-based biases have been studied (e.g., past work showed that if we edit an image of a woman with a prompt like “present the person as a senior executive”, models often edit the woman into a man) [60]. We discussed related works in detail in Section A.

2 Evaluation Setup

2.1 Problem Statement

We focus on multi-turn editing, where a model iteratively edits an image with prompts (or instructions) in sequence. We begin with an input image I_0 (also called the base image) and a prompt P_1 , which is called an editing instruction. A text-guided image editing model

M takes the prompt P_1 and edits the base image I_0 to produce an edited image I_1 ; $I_1 = M(I_0)$. Next, the model takes the edited image I_1 and applies a new editing instruction P_2 to obtain I_2 . This process continues for k editing instructions, resulting in the final edited image I_K . This iterative sequence of edits from I_0 to I_K is called an edit session. In this work, $K = 4$ (chosen based on computational constraints). Figure 1 shows an example edit session.

2.2 Face Datasets (Base Images)

To obtain the base images (I_0), we used two popular celebrity face datasets: CelebA: a dataset with over 200k images covering 10k celebrities [49] and CelebSET [57]: over 12k images covering 80 celebrities (we used face images from CelebSET’s source dataset: IMDB-WIKI [59]). In both CelebA and CelebSET, each image contains a headshot of a celebrity. We removed images with multiple faces. CelebA provides binary gender (male, female) and age annotations (young, old) for each celebrity. CelebSET provides annotations for gender (male, female), skin tone (light, dark), and age (numerical value). Using age values, we created two categories: young (faces in the age range 20 – 30) and old (faces in the age range 50 – 60). We only considered these age ranges to obtain two distinct age groups.

Due to the computational cost of text-guided models, we limited our analysis to ~ 500 images per dataset. To filter images from each dataset, we used a facial similarity score (FSS) threshold (refer to Section 2.5.1 for the FSS metric). We used $FSS > 0.95$ for CelebA and > 0.85 for CelebSET. This resulted in 549 images for CelebA and 517 images for CelebSET.

CelebA data statistics for gender: 244 images from 208 males and 305 images from 280 females. Age statistics: 413 images from 375 young people and 136 images from 116 old people.

CelebSET data statistics for gender-race: 222 images from 20 light-skinned males and 95 images from 16 dark-skinned males; 161 images from 17 light-skinned females and 39 images from 15 dark-skinned females. Age statistics: 98 images from 28 young people (aged 20 – 30) and 41 images from 8 old people (aged 50 – 60).

2.3 Editing Attributes

To generate editing instructions, we manually curated Face-Edit-Attributes, a dataset of 169 facial editing attributes (e.g., stetson hat). We designed a hierarchical approach to curate this dataset (Figure 2). At level 1, we defined three broad editing types (or axes): Hair, Accessories, and Pose. Each axis is divided into sub-axes (level 2). For example, in the axis “Hair”, we defined the following sub-axes: hair style, hair color, hair type, mustache, and beard.

Each sub-axis has multiple categories (level 3). For instance, within the sub-axis “hair style”, we have categories like buzz cut, bob cut, and spiked cut. Finally, each category has multiple editing attributes (level 4) closely related to the category. For instance, within the category “spiked cut”, we have editing attributes like spiked hair, liberty spiked hair, mohawk hair, and gelled spiked hair. The editing attributes are used to create the editing instructions. Refer to Section B for a complete list of editing attributes.

This hierarchical design has several advantages: (1) We can create a coarse-grained test set (e.g., Hair-Test), as well as a fine-grained test set (e.g., Mustache-Test). (2) This design forms the basis of our new overediting metric, which we discuss in Section 2.5.3.

We chose three axes (hair, accessories, and pose) to purely focus on the facial region. For the hair axis, we chose sub-axes focused on scalp and facial hair. For accessories, we chose glasses and hats as our sub-axes. We ignored non-facial sub-axes like hand or neck accessories. We did not consider earrings: if a woman’s ears are occluded in the base image, adding an earring would require the model to generate an ear (or estimate ear position). This adds a layer of complexity to the model that can impact editing results. For pose, we used a single sub-axis focused on head/gaze movements.

In each sub-axis, we chose categories (level 3) with a commonly accepted shape and appearance definition. We excluded several categories that have a fluid shape definition. Consider the glass sub-axis. Glasses have several popular categories: aviators, cat-eyed, geometric, and browline. In most cases, these categories look similar, as each glass brand follows its own definition. This affects editing alignment, since we cannot be sure if the model executed the edit correctly. For the head/gaze sub-axis, we focused on three categories: looking left, right, and straight ahead. We excluded looking up and down movements as this can impact identity preservation, which is integral to our study.

In each category, we chose attributes (level 4) that encourage semantic and linguistic diversity in editing instructions. We introduced semantic diversity through synonyms (e.g., snapback cap for baseball hat). For linguistic diversity, we used alternative phrasings (e.g., looking ahead, frontal view for “looking straight ahead”).

There are other editing attributes relevant to facial edits that we did not consider. We excluded attributes that modify the face (e.g., jawline and nose modifications) as this affects identity preservation. We excluded subtle edits such as facial expressions and makeup. Facial expressions are subtle and hard to evaluate (e.g., it is hard to distinguish a surprised reaction from a sad one). Makeup edits are hard to identify as the image hue can obscure makeup shades.

2.4 Editing Instructions and Sessions

To generate an editing instruction, we selected an attribute from the Face-Edit-Attributes dataset. For instance, to evaluate hair-based edits, we randomly select an attribute (say crimson red hair) from the Hair axis. We then prompt a multimodal large language model (MLLM) to generate an editing instruction (e.g., “make the person’s hair color crimson red”) by combining the attribute with an appropriate verb. We used Gemini-2.5-Flash [12] as our MLLM.

To create an edit session, the MLLM generates 4 instructions in sequence (since each session has 4 edits). After generating the first instruction, P_1 (e.g., “make the person’s hair color crimson red”), we provided the MLLM with metadata about P_1 (attribute: crimson red hair, sub-axis: hair-color, axis: hair) to give it context to generate the next instruction P_2 . We provided our prompt in Section C.

We created four evaluation sets: one set for each axis (Hair-Test, Accessories-Test, Pose-Test) and an overall set (Multi-Axis-Test). To create the evaluation set for a specific axis, we randomly selected editing attributes only from that axis. For the Multi-Axis-Test set, we randomly selected attributes from the complete set of 169 attributes. For Hair-Test, Accessories-Test, and Multi-Axis-Test, we generated 10 edit sessions per test set. For Pose-test, we generated 5 sessions (since poses have only three variations: look left, look straight, and look right). This resulted in a total of 35 edit sessions or 140 editing

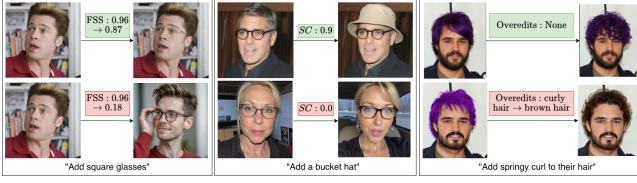


Figure 3: Visually comparing good and bad scenarios for our evaluation metrics. Left: the top row (Nano-Banana) shows good identity preservation, while the bottom row (Flux-Dev) shows poor identity preservation. Center: the top row (Nano Banana) shows a correctly executed edit, while the bottom row (Flux-Dev) shows a poorly executed edit. Right: the top row (Nano Banana) shows a case for no overedits, while the bottom row (Qwen-Edit) indicates an overedit.

instructions (4 instructions per session). Refer to Section D for the list of edit sessions in each test set.

We generated simple, one-line editing instructions consistent with the instruction template used by popular general scene editing benchmarks [16, 63, 71]. These instructions test the effectiveness of text-guided models for real-world facial editing applications and enable larger-scale evaluations than detailed editing instructions.

2.5 Evaluation Metrics

2.5.1 Identity preservation. We verified whether the facial identity of a person is preserved across the edit session using a Facial Similarity Score (FSS), which ranges from 0 to 1. To compute FSS, we used Buffalo-Large [1], a popular face recognition model, which uses RetinaFace [26] to detect a face in the image, and a ResNet-50 [29] network to generate an embedding vector for the detected face.

Assume we have an image I_k (k is the editing step) of a celebrity C . We computed an embedding vector $f(I_k)$. For celebrity C , we loaded a set of reference images R_C from our reference database (the reference database has images for each celebrity, which are extracted from CelebA and CelebSET; they are not part of the base images we used for editing). We computed an embedding vector $f(r)$ for each reference image $r \in R_C$. We computed cosine similarity between $f(I_k)$ and each $f(r)$ from the reference set. FSS_k is the maximum of these similarity scores. We computed FSS in each edit session, from I_0 to the final edited image I_k . An FSS_k close to 1 indicates that the person in I_k closely resembles celebrity C . Figure 3 (left) shows examples of edits with high and low FSS.

To compare model performance in an edit session, we also propose Drop_{FSS} , which is an extension to FSS. In Drop_{FSS} , we calculated the average % drop in FSS of a model across an edit session.

$$\text{Drop}_{\text{FSS}} = \frac{1}{K} \sum_{k=1}^K \left(\frac{FSS_0 - FSS_k}{FSS_0} \right) \times 100\%.$$

$K = 4$ because we have four edits per session. FSS_0 is the facial similarity score for I_0 (base image), and FSS_k is the score for I_k . A Drop_{FSS} of 40% indicates that across an edit session, the FSS dropped by 40% (on average) with respect to the base image. A lower Drop_{FSS} indicates better identity preservation.

2.5.2 Editing Alignment. To measure how well a model followed the editing instructions, we used Semantic Consistency (SC), a popular MLLM-based metric proposed by [41] to measure alignment in scene editing benchmarks. The MLLM (we used Gemini-2.5-Flash)

outputs two scores: (1) editing success: 0 indicates the editing instruction was not executed at all, while 10 indicates perfect execution, (2) overediting degree: 0 indicates that the image was overedited (e.g., unwanted background changes), and 10 indicates that the image showed minimal changes. For a strict scoring, we reported SC as the minimum of these two scores. We normalized SC in the range of 0 – 1. We computed SC on images in $[I_1, I_k]$, as SC can only be computed for an edited image. Refer to Figure 3 (center) for examples of edits with high and low SC.

The overediting score in SC has limitations: (1) In many cases, it does not capture overedits correctly. For example, in Figure 3, the bottom row of the rightmost grid shows an overedit: the model changed hair color when it was prompted to change hair style. In this example, $SC = 0.8$, which means that both the “editing success” and “overediting degree” scores were high. While editing success should be high, the overediting score should be low. (2) It cannot measure how many overedits occurred in an image and what these overedits were (this is critical to assess spurious correlations). To provide a more accurate and detailed measure of overediting, we propose a new metric called the Overediting Matrix.

2.5.3 Overediting Matrix. To count overedits in a model, we propose an Overediting Matrix. This is a 35×35 matrix where the rows and columns represent 35 categories of the Face-Edit-Attributes dataset (level 3). Each row represents the category that the user intended to edit, and each column represents the category that the model edited in the image. For example, if the matrix entry [red hair][stubble] equals 645, it means that in 645 instances, the prompt requested a change in hair color to red, but the model added stubble.

To create this matrix, we used a state dictionary. Assume we have a base image (I_0) of a person with black-colored straight hair, no facial hair, looking ahead. For I_0 , the state dictionary will be: {hair-color: black, beard: clean shaven, hair-type: straight, head position: looking forward, hair-style: None, mustache: None, hats: None, glasses: None}. The keys in this dictionary are the 8 sub-axes from the Face-Edit-Attributes dataset. The values are the categories in these sub-axes (e.g., red hair, blue hair).

Suppose a model edited I_0 using the instruction P_1 : “make the person’s hair color crimson red”. Assume the model changed the hair color to crimson red, but also added stubble (overedit). The updated dictionary would be: {hair-color: black $\rightarrow$ red, beard: clean shaven $\rightarrow$ stubble, ... } (other sub-axes stay unchanged). The hair color changing from black $\rightarrow$ red is an intended edit. However, beard changing from clean shaven $\rightarrow$ stubble is an overedit. In our overediting matrix, we increment [red hair][stubble] by 1.

To track the state dictionary after each edit, we used Gemini-2.5-Flash. Given an image I_0 , the MLLM predicts the most probable category for each sub-axis. It returns None for sub-axes not present in the image. If the MLLM selects multiple top categories for a sub-axis (e.g., black hair, brown hair), we record all of them.

We selected Gemini-2.5-Flash as our state tracker as it showed the highest alignment with human annotators. We compared state dictionaries generated by four models (Gemini-2.5-Flash [12], Gemini-2.5-Pro [13], GPT-4o [33], and CLIP [56]) against the state dictionary created by human annotators. Alignment accuracies: 79.65% for Gemini-2.5-Flash, 79.26% for Gemini-2.5-Pro, 70.31% for GPT-4o, and 35.47% for CLIP. Refer to Section E for details.

We tracked coarse-grained categories rather than the 169 fine-grained attributes, as MLLMs more reliably distinguish categories (e.g., red vs. blue hair) than closely related attributes (e.g., crimson red vs. ruby red hair). Coarse-grained categories are sufficient for counting overedits.

To compare the amount of overediting across models, we also computed an overediting rate (OER) using the state dictionary. In a session (of four edits), if a model performs a total of two overedits, the OER of that session is $2/4$ or 0.5 .

2.5.4 Wasserstein Distance. To quantify performance disparities across demographic groups, we used Wasserstein Distance (WD), a popular fairness metric [51]. WD measures the distance between two distributions. Consider two distributions: (1) Drop_{FSS} values for male faces, and (2) Drop_{FSS} for female faces. If $WD = 2$, this indicates that the male and female Drop_{FSS} values are 2% apart. WD does not give the direction of bias (here, we do not know if Drop_{FSS} is higher for males or females). To get the direction of bias, we must examine the mean values of the demographic groups. WD ranges from 0 to the largest value of the chosen metric.

For each of the following metrics: Drop_{FSS} , SC, and OER, we reported three WD scores: WD_g (gender disparity between male and female groups), WD_s (skin-tone disparity between light and dark groups), and WD_{age} (age disparity between young and old groups). To assess whether WD values are statistically significant, we reported the p-values. Our null hypothesis was that two groups (e.g., male and female) are drawn from the same distribution (no disparity). We tested our null hypothesis using the standard Monte Carlo permutation test. If $p < 0.05$, we reject the null hypothesis and say that the WD score is statistically significant.

2.6 Editing Models

We evaluated five popular text-guided image editing models on CelebA: Flux.1-Dev (Flux-Dev) [5], Flux.1 Kontext-Pro (Kontext-Pro) [43], Gemini-2.5-Flash-Image (Nano-Banana) [11], Bagel-Edit [25], and Qwen-Edit [65]. On CelebSET, we also evaluated SeedEdit 3.0 [62]. We limited SeedEdit 3.0 to a single dataset due to cost constraints. If a model did not generate all four edited images in an edit session, we excluded that session. Across all models, we evaluated about 200K sessions ($\sim 1\text{M}$ images) across both datasets. Refer to Section F for the number of edit sessions per model.

3 Results

3.0.1 Overall comparisons. We compared identity preservation across test sets. In Table 1, for each model, we computed a mean Drop_{FSS} on each test set by averaging the Drop_{FSS} percentages across all sessions in the test set. A low Drop_{FSS} is better, as it indicates that the model better preserved facial identity.

For most models, FSS dropped in the range of 22 - 32% across four edits. In Table 1, (Avg. column), for CelebA, all models (except Flux-Dev) showed average Drop_{FSS} in the range of 28.7 – 32.7%. In CelebSET, we saw similar average Drop_{FSS} in 4 out of 6 models (22.3 – 30.8%). Across CelebA and CelebSET, most models best preserved facial identity for hair edits (blue numbers), and worst preserved identity for pose edits (red numbers).

Table 1: Drop_{FSS} (lower is better) across all test sets on CelebA and CelebSET. For each model, blue number indicates its best performance, while red indicates its worst performance across the four test sets. The Avg. column shows the average performance of a model across test sets. The bolded number indicates the overall best-performing model. Subscripts denote 95% confidence intervals.

CelebA					
Model	Hair↓	Acc.↓	Pose↓	Multi↓	Avg.↓
Kontext-Pro	24.60 _{0.16}	25.37 _{0.14}	37.48 _{0.28}	32.57 _{0.17}	28.68 _{0.10}
Nano-Banana	23.08 _{0.12}	33.51 _{0.14}	32.43 _{0.15}	32.06 _{0.16}	29.93 _{0.08}
Bagel-Edit	18.79 _{0.09}	27.37 _{0.10}	52.98 _{0.28}	33.51 _{0.20}	30.33 _{0.11}
Qwen-Edit	28.04 _{0.20}	27.90 _{0.14}	47.59 _{0.22}	34.82 _{0.17}	32.73 _{0.10}
Flux-Dev	89.13 _{0.06}	89.58 _{0.06}	89.68 _{0.09}	89.40 _{0.06}	89.41 _{0.04}
CelebSET					
Model	Hair↓	Acc.↓	Pose↓	Multi↓	Avg.↓
Kontext-Pro	21.47 _{0.16}	26.58 _{0.12}	29.39 _{0.30}	29.32 _{0.18}	26.21 _{0.09}
Nano-Banana	16.56 _{0.10}	25.97 _{0.10}	23.02 _{0.16}	23.85 _{0.14}	22.25 _{0.06}
Bagel-Edit	18.20 _{0.10}	27.73 _{0.09}	56.25 _{0.29}	33.79 _{0.22}	30.81 _{0.12}
Qwen-Edit	46.56 _{0.24}	46.19 _{0.22}	39.40 _{0.23}	44.40 _{0.23}	44.82 _{0.12}
Flux-Dev	83.53 _{0.07}	83.96 _{0.07}	83.58 _{0.11}	83.65 _{0.07}	83.69 _{0.04}
SeedEdit 3.0	18.64 _{0.11}	25.50 _{0.10}	45.54 _{0.20}	31.68 _{0.14}	28.15 _{0.09}

Table 2: SC (higher is better) across all test sets on CelebA and CelebSET. For each model, blue number indicates its best performance, while red number indicates its worst performance across the four test sets. In the Avg. column, the bolded number indicates the best-performing model. Values scaled by 100.

CelebA					
Model	Hair↑	Acc.↑	Pose↑	Multi↑	Avg.↑
Kontext-Pro	76.10 _{0.50}	73.40 _{0.60}	47.40 _{0.90}	64.00 _{0.50}	68.40 _{0.30}
Nano Banana	87.90 _{0.30}	84.70 _{0.40}	62.60 _{0.80}	75.10 _{0.50}	79.60 _{0.30}
Bagel-Edit	80.30 _{0.40}	87.80 _{0.30}	59.60 _{0.70}	68.90 _{0.50}	76.20 _{0.30}
Qwen-Edit	76.60 _{0.50}	79.90 _{0.40}	67.50 _{0.80}	64.70 _{0.50}	72.90 _{0.30}
Flux-Dev	30.30 _{0.50}	33.70 _{0.50}	31.40 _{0.80}	30.20 _{0.50}	31.40 _{0.30}
CelebSET					
Model	Hair↑	Acc.↑	Pose↑	Multi↑	Avg.↑
Kontext-Pro	75.60 _{0.40}	73.10 _{0.60}	37.00 _{1.00}	61.20 _{0.60}	66.10 _{0.40}
Nano Banana	85.90 _{0.30}	82.30 _{0.40}	59.40 _{0.80}	73.00 _{0.60}	77.40 _{0.30}
Bagel-Edit	78.30 _{0.40}	86.80 _{0.40}	60.50 _{0.70}	67.70 _{0.60}	75.10 _{0.30}
Qwen-Edit	78.90 _{0.40}	82.50 _{0.40}	63.00 _{0.80}	66.30 _{0.50}	74.00 _{0.30}
Flux-Dev	28.80 _{0.50}	28.70 _{0.50}	29.30 _{0.80}	28.20 _{0.50}	28.70 _{0.30}
SeedEdit 3.0	82.10 _{0.40}	86.30 _{0.40}	63.80 _{0.80}	68.90 _{0.50}	77.00 _{0.30}

Nano-Banana and Kontext-Pro performed best at preserving facial identity, while Flux-Dev performed the worst. In Table 1, for CelebA, Kontext-Pro showed the best aggregated Drop_{FSS} of 28.68%. For CelebSET, Nano-Banana showed the best aggregated Drop_{FSS} of 22.25%. For Flux-Dev, the aggregated Drop_{FSS} was very high: $> 80\%$ in both CelebA and CelebSET.

To compare editing alignment, for each model, we computed an average SC score per session by averaging the SC scores across the edits in that session. In Table 2, for each model, we calculated the mean of the averaged SC scores for all sessions in each test set. We scaled the SC values by 100. A higher SC is better, as it indicates that the model correctly executed the editing instruction.

Table 3: Frequent overediting patterns across models. Nano-Banana (NB), SeedEdit (SE), Bagel-Edit (BE), Qwen-Edit (QE), Flux-Dev (F-Dev), Kontext-Pro (K-Pro).

CelebA	
Intended Edit → Overedit	Models
bob cut → straight hair	BE, K-Pro, QE
sideburn mutton chops → handlebar mustache	BE, K-Pro, QE
bob cut → brown hair	F-Dev, QE
sideburn mutton chops → pencil mustache	K-Pro, QE
straight hair → brown hair	BE, QE
CelebSET	
Intended Edit → Overedit	Models
sideburn mutton chops → handlebar mustache	BE, K-Pro, QE, SE
bob cut → brown hair	F-Dev, K-Pro, QE
curly hair → None	BE, NB, SE
bob cut → straight hair	K-Pro, QE
curly hair → brown hair	F-Dev, QE

Table 4: OER (lower is better) across all test sets on CelebA and CelebSet. For each model, blue number indicates its best performance, while red number indicates its worst performance across the test sets. The bolded number indicates the best-performing model. Values scaled by 100.

CelebA					
Model	Hair↓	Acc.↓	Pose↓	Multi↓	Avg.↓
Kontext-Pro	19.54 _{0.67}	11.38 _{0.64}	16.47 _{1.10}	22.83 _{0.77}	17.75 _{0.39}
Nano Banana	15.02 _{0.59}	10.42 _{0.71}	10.18 _{0.83}	17.62 _{0.69}	13.79 _{0.35}
Bagel-Edit	18.47 _{0.64}	8.48 _{0.56}	13.82 _{0.95}	21.03 _{0.73}	15.68 _{0.35}
Qwen-Edit	24.87 _{0.76}	13.66 _{0.74}	19.87 _{1.05}	27.25 _{0.90}	21.63 _{0.43}
Flux-Dev	66.42 _{1.33}	68.38 _{1.31}	47.76 _{1.58}	65.31 _{1.29}	64.00 _{0.70}
CelebSET					
Model	Hair↓	Acc.↓	Pose↓	Multi↓	Avg.↓
Kontext-Pro	23.53 _{0.74}	15.63 _{0.81}	14.95 _{1.14}	25.90 _{0.84}	20.90 _{0.44}
Nano Banana	16.99 _{0.63}	21.72 _{1.41}	9.71 _{0.81}	24.15 _{1.05}	19.34 _{0.57}
Bagel-Edit	24.10 _{0.88}	12.77 _{0.99}	13.81 _{0.94}	28.36 _{1.08}	20.61 _{0.51}
Qwen-Edit	32.08 _{0.84}	22.88 _{0.92}	18.64 _{1.05}	33.17 _{0.97}	27.85 _{0.49}
Flux-Dev	72.23 _{1.40}	68.07 _{1.37}	52.13 _{1.73}	72.95 _{1.46}	68.38 _{0.75}
SeedEdit 3.0	18.84 _{0.69}	9.29 _{0.62}	18.34 _{1.08}	25.12 _{1.04}	17.80 _{0.43}

Editing models best execute accessory and hair edits, and worst execute pose edits. In Table 2, for CelebA and CelebSET, models showed the best (highest) SC either on accessory or on hair edits (blue numbers). Across CelebA and CelebSET, most models reported their worst SC on pose-edits (red numbers).

Nano-Banana and Seed-Edit are the best models to execute edits, while Flux-Dev is the worst. In Table 2, for CelebA, Nano-Banana showed the best average SC at 79.6. For CelebSET, both Nano-Banana and Seed-Edit showed the best average SC at 77. Flux-Dev SC was much lower than other models in both datasets.

We examined overediting patterns in models. For each model, across all test sets, we computed one Overediting Matrix each for CelebA and CelebSET. We used this matrix to identify frequent overedits. For example, assume we ask a model to add a “handlebar mustache” to 100 images. For 15 or more of these images, say our model did the same overedit — changing the hairstyle to curly hair. In this case, we call “handlebar mustache → curly hair” as an overedit pattern. In Table 3, we showed the most common overedit patterns for all models in CelebA and CelebSET.

Models frequently made overedits when performing hair-based edits. In Table 3, for CelebA and CelebSET, we observed common overedit patterns for hair-based edits (e.g. mutton chops → handlebar mustache, bob cut → straight hair). These overedit patterns were prevalent in Bagel-Edit, Kontext-Pro, and Qwen-Edit.

To compare the amount of overediting across models, we used the OER metric (Section 2.5.3). In Table 4, for each model, we reported the mean OER for a test set by averaging the OER values across all sessions in that set. We scaled OER values by 100. A lower OER is better, as it indicates fewer overedits.

Editing models created the fewest overedits for accessory or pose edits, and the most overedits for multi-axis edits. In Table 4, for CelebA and CelebSET, almost all models achieved their best (lowest) OER either on accessory or on pose edits (blue numbers). Almost all models achieved their worst OER on multi-axis edits (red numbers).

Nano-Banana and SeedEdit created the fewest overedits in a session. In Table 4 (Avg. column), for CelebA, Nano-Banana showed the lowest average OER at 13.8. For CelebSET, SeedEdit showed the lowest average OER at 17.49.

3.0.2 Demographic comparisons. We examined demographic biases in identity preservation using Wasserstein Distance or WD (Section 2.5.4). If a model has $WD > 3$ (and $p < 0.05$), we consider the model to have a prominent bias (refer to Section G to understand why we chose $WD > 3$). In Table 5, we reported WD_g (gender disparity between male and female $Drop_{FSS}$ values) for CelebA. We also reported mean values for both gender groups. In the “mean” column, the left value is the mean of the worst performing group, and the right value is the mean of the best group.

For CelebSET (Table 6), we reported WD_g and WD_s (skin-tone disparity between light and dark groups). CelebSET has four intersectional groups: male-light (ML), male-dark (MD), female-light (FL), and female-dark (FD). For conciseness, we only reported the mean values of the two worst intersectional groups. In the “mean” column, the left value is the mean of the worst-performing group, and the right value is the mean of the second-worst group. The two-worst mean values, combined with WD_g and WD_s are sufficient to examine bias trends.

Qwen-Edit is poor at retaining the facial identity of females. In CelebA (Table 5), for Qwen-Edit we observed a gender bias for pose-edits ($WD_g = 7.3$). The worst performing group was female (F). In CelebSET (Table 6), we observed a gender bias in Qwen-Edit for all test sets. For each test set, the worst two groups were females (FL and FD).

Flux-dev is poor at retaining the facial identity of people with darker skin tones. In CelebSET (Table 6), for Flux-Dev, we observed skin-tone bias in all test sets ($WD_s > 3$). The two worst groups were always dark-skinned people (FD and MD).

We showed age bias results for identity preservation in Section H. With CelebA, we did not observe clear age bias trends (we saw age bias in a few models only for pose-edits). With CelebSET, we observed age bias across all test sets. Most models (except Qwen-Edit) showed worse $Drop_{FSS}$ on old faces.

To examine demographic bias in editing alignment, we computed WD scores using SC values. We used the same reporting format as

Table 5: Drop_{FSS} % (lower is better) disparity on CelebA. For each model and test set, we reported mean Drop_{FSS} for Male (M) and Female (F). In each mean column, the left value is the mean of the worst performing group, and the right value is the mean of the best group. WD_g denotes the Wasserstein distance between the male and female Drop_{FSS} distributions. A superscript * marks non-significant values ($p > 0.05$). $WD > 3$ is highlighted as it indicates gender bias.

Model	Accessories↓			Hair↓			Pose↓			Multi-axis↓		
	Mean		WD_g	Mean		WD_g	Mean		WD_g	Mean		WD_g
Flux-Dev	F:89.7	M:89.4	0.4	M:89.2	F:89.1	0.4	F:90.0	M:89.2	0.8	F:89.5	M:89.3	0.5
Kontext-Pro	M:25.8	F:25.0	1.6	F:24.7	M:24.4	1.2	F:39.1	M:35.5	3.4	F:33.6	M:31.2	2.4
Nano Banana	M:34.5	F:32.8	1.8	M:24.3	F:22.1	2.3	M:33.0	F:32.0	1.0	M:33.0	F:31.3	2.2
Bagel-Edit	M:28.1	F:26.8	1.4	M:19.5	F:18.3	1.2	F:55.2	M:50.2	5.0	F:33.8	M:33.1	1.4
Qwen-Edit	M:29.1	F:27.0	2.5	F:28.8	M:27.1	2.0	F:50.8	M:43.6	7.3	F:35.3	M:34.2	1.8

Table 6: Drop_{FSS} % (lower is better) disparity on CelebSET. We reported the mean Drop_{FSS} for the two worst intersectional groups. In each mean column, left value: worst; right value: second-worst mean group. WD_g and WD_s denote Wasserstein distances between gender and skin-tone groups, respectively. A superscript * marks non-significant values ($p > 0.05$). ML: Male-Light, MD: Male-Dark, FL: Female-Light, FD: Female-Dark. $WD > 3$ is highlighted.

Model	Accessories↓			Hair↓			Pose↓			Multi-axis↓		
	Mean: 2 worst groups		WD_g	Mean: 2 worst groups		WD_g	Mean: 2 worst groups		WD_g	Mean: 2 worst groups		WD_g
Flux-Dev	FD:89.6	MD:85.5	0.9	FD:88.7	MD:85.8	0.5	FD:89.9	MD:85.2	1.3	FD:89.3	MD:85.1	0.8
Kontext-Pro	ML:28.1	FL:26.3	2.2	FL:22.0	ML:21.7	1.7	FD:33.5	FL:32.6	5.6	FL:31.0	ML:29.0	2.0
Nano-Banana	ML:27.2	MD:26.0	2.2	MD:17.5	FL:16.8	0.4*	MD:25.1	FD:23.8	0.7*	FL:24.1	ML:24.0	1.6
Bagel-Edit	ML:29.2	MD:27.2	2.4	MD:19.4	FD:18.8	1.9	MD:58.0	FD:57.6	1.2*	MD:34.8	ML:33.8	1.3
Qwen-Edit	FL:50.6	FD:47.3	6.4	FL:52.0	FD:47.8	7.6	FD:45.4	FL:43.8	7.7	FL:49.8	FD:45.3	7.4
SeedEdit 3.0	ML:27.3	MD:25.7	3.5	MD:20.6	ML:19.2	3.1	FL:46.0	MD:46.0	0.5*	MD:32.3	ML:31.9	2.0

Table 7: SC (higher is better) disparity on CelebA. We reported mean SC values for Male (M) and Female (F). In each mean column, left: worst, right: best group. WD_g denotes the Wasserstein distance between the male and female SC distributions. A superscript * marks non-significant values ($p > 0.05$). $WD > 3$ is highlighted. Values are scaled by 100.

Model	Accessories↑			Hair↑			Pose↑			Multi-axis↑		
	Mean		WD_g	Mean		WD_g	Mean		WD_g	Mean		WD_g
Flux-Dev	M:33.2	F:34.1	0.9*	F:28.2	M:32.9	4.6	F:30.1	M:33.0	3.0	F:29.0	M:31.8	2.9
Kontext-Pro	F:73.4	M:73.5	1.8	M:75.5	F:76.7	2.0	F:46.5	M:48.6	2.6	F:63.5	M:64.6	1.2
Nano Banana	M:83.7	F:85.6	1.9	M:87.0	F:88.6	1.7	F:61.9	M:63.4	1.8	M:74.0	F:75.9	2.0
Bagel-Edit	M:87.5	F:88.0	0.6	M:79.0	F:81.2	2.3	M:58.8	F:60.2	1.6	M:68.8	F:69.0	1.4
Qwen-Edit	M:78.6	F:80.9	2.3	F:76.4	M:76.9	1.1	M:66.1	F:68.7	2.6	M:63.4	F:65.8	2.5

Table 8: SC (higher is better) disparity on CelebSET. We reported the mean SC values for the two worst intersectional groups. In each mean column, left: worst, right: second-worst group. WD_g and WD_s denote Wasserstein distances between gender and skin-tone groups, respectively. * marks non-significant values ($p > 0.05$). $WD > 3$ is highlighted. Values scaled by 100.

Model	Accessories↑			Hair↑			Pose↑			Multi-axis↑		
	Mean: 2 worst groups		WD_g	Mean: 2 worst groups		WD_g	Mean: 2 worst groups		WD_g	Mean: 2 worst groups		WD_g
Flux-Dev	ML:28.1	FL:28.9	0.8*	FL:25.1	FD:26.6	5.5	FD:26.2	FL:27.8	2.9	FL:25.3	FD:27.3	4.1
Kontext-Pro	FL:72.2	ML:73.5	2.8	FL:74.6	ML:75.2	0.6*	MD:35.1	FD:35.4	3.3	FD:60.4	FL:60.5	1.2*
Nano Banana	FD:74.8	MD:80.9	1.5	ML:84.2	FL:86.8	2.4	FD:58.2	FL:58.8	2.1	MD:71.6	FL:73.1	1.1*
Bagel-Edit	FL:85.7	FD:86.9	1.5	MD:76.7	FL:77.8	1.3	FL:59.6	FD:60.8	1.1*	ML:67.2	FL:67.5	0.7*
Qwen-Edit	ML:82.0	FL:82.4	1.2	FL:76.2	ML:79.7	2.9	ML:60.7	MD:61.7	5.1	ML:64.9	FL:66.2	1.3
SeedEdit 3.0	FL:85.7	FD:86.0	1.0	ML:80.9	MD:82.1	2.2	FD:62.5	ML:63.5	0.9*	FD:67.1	MD:68.2	0.8*

Drop_{FSS}. CelebA results are shown in Table 7, and CelebSET results are shown in Table 8 (for SC, lower score is worse).

When executing edits, most models showed marginal (or no) gender and skin-tone bias for most edit types. In most models, we did not observe clear trends for gender and skin-tone bias across both datasets (Tables 7 and 8). With Flux-Dev, we observed gender bias ($WD_g > 3$) in some test sets across both datasets. In these cases, the worst performing group was always female.

We showed our age bias analysis for editing alignment in Section H. On CelebA, only Flux-Dev exhibited age bias, with lower SC for older faces. On CelebSET, we observed no age bias trends.

To examine demographic bias in overediting, we computed WD using OER values. CelebA results are shown in Table 9, and CelebSET results are shown in Table 10.

Models created more overedits when editing males, especially dark-skinned males. In CelebA (Table 9) across all test sets, we observed a significant gender bias for all models. The worst group (highest OER) was Male (M). In CelebSET (Table 10), across test sets, we observed significant gender or skin-tone bias. In all models, the worst group was dark-skinned males (MD).

In Section H, we showed age bias results for overediting. For CelebA and CelebSET, all models showed age bias in all test sets. In most cases, models showed worse OER on old faces. For all our demographic bias analyses (with Drop_{FSS}, SC, and OER), we provided mean values for all CelebA and CelebSET groups (with confidence intervals) in Section I.

We conducted an additional analysis to examine which specific editing attributes commonly led to demographic biases in overediting. Refer to our analysis in Section J.

Table 9: OER (lower is better) disparity on CelebA. We reported mean OER values for Male (M) and Female (F). In each mean column, left: worst, right: best group. WD_g denotes the Wasserstein distance between the male and female OER distributions. * marks non-significant values ($p > 0.05$). $WD > 3$ is highlighted. Values are scaled by 100.

Model	Accessories↓			Hair↓			Pose↓			Multi-axis↓		
	Mean	Values	WD_g	Mean	Values	WD_g	Mean	Values	WD_g	Mean	Values	WD_g
Flux-Dev	M:82.5	F:57.1	25.3	M:85.0	F:51.6	33.5	M:65.3	F:33.8	31.5	M:84.4	F:50.0	34.4
Kontext-Pro	M:14.9	F:8.6	6.3	M:25.5	F:14.7	10.8	M:21.9	F:12.1	10.0	M:28.3	F:18.5	9.8
Nano Banana	M:15.0	F:6.8	8.3	M:18.0	F:12.6	5.3	F:10.5	M:9.8	1.0*	M:21.9	F:14.2	7.7
Bagel-Edit	M:11.7	F:6.0	5.9	M:23.5	F:14.4	9.1	M:16.1	F:12.0	4.5	M:26.9	F:16.3	10.6
Qwen-Edit	M:16.2	F:11.7	4.5	M:30.4	F:20.4	10.0	M:26.4	F:14.7	11.7	M:33.3	F:22.4	11.1

Table 10: OER (lower is better) disparity on CelebSET. We reported the mean OER values for the two worst intersectional groups. In each mean column, left: worst, right: second-worst group. WD_g and WD_s denote Wasserstein distances between gender and skin-tone groups, respectively. * marks non-significant values ($p > 0.05$). $WD > 3$ is highlighted. Values scaled by 100.

Model	Accessories↓				Hair↓				Pose↓				Multi-axis↓			
	Mean: 2 worst groups		WD_g	WD_s	Mean: 2 worst groups		WD_g	WD_s	Mean: 2 worst groups		WD_g	WD_s	Mean: 2 worst groups		WD_g	WD_s
Flux-Dev	MD:93.7	ML:69.9	23.1	19.9	MD:106.7	ML:76.9	35.1	24.5	MD:84.5	ML:56.0	32.4	24.4	MD:101.3	ML:81.1	36.8	22.3
Kontext-Pro	MD:25.5	ML:16.4	8.9	7.1	MD:38.5	FD:20.9	5.7	13.3	MD:19.0	ML:16.6	6.4	2.3*	MD:41.4	ML:24.3	9.2	13.1
Nano Banana	FD:50.3	MD:32.9	8.5	22.0	MD:32.1	FD:15.3	6.6	13.8	MD:15.6	ML:9.2	3.9	3.7	MD:39.0	ML:21.5	9.8	11.4
Bagel-Edit	MD:19.9	FL:15.0	4.9	4.7	MD:38.7	FL:23.2	7.1	13.2	MD:17.3	ML:14.4	3.8	1.4*	MD:42.8	ML:29.1	12.5	11.5
Qwen-Edit	MD:36.2	ML:22.2	9.0	9.2	MD:49.9	FL:28.5	6.5	14.8	MD:25.3	ML:21.5	10.3	2.4*	MD:52.1	ML:31.6	11.8	15.3
SeedEdit 3.0	MD:17.8	ML:8.1	4.5	6.7	MD:33.3	ML:17.0	8.0	12.4	MD:23.5	ML:18.7	4.7	2.6	MD:46.4	FD:35.0	6.4	24.5

To ensure our bias insights are reliable, we conducted additional experiments: (1) In Section K, we evaluated if FSS showed demographic bias in base images. We found no gender, age, or skin tone bias. (2) In Section L, we conducted an experiment to check if FSS is sensitive to occlusions (if a face in the base image is occluded by glasses or hats, does FSS drop w.r.t a clean face?). We found that FSS was similar for faces with or without occlusions. These experiments proved that any bias in identity preservation came from the editing models. (3) In Section M, we conducted a controlled experiment to check if text-guided models were biased toward stereotypical edits (do models perform better on males if instructions are male-centric? — e.g., add a handlebar mustache). We found that models do not show such stereotypical biases.

4 Discussion

Our results show that Nano-Banana is the best model for facial editing and Flux-Dev is the worst. While Flux-Dev supports image editing, it was optimized for image generation. Flux-Dev often produces cartoonish edits, resulting in poor evaluation scores.

Models are reliable for short facial editing sequences (3 – 4 edits). We do not recommend using these models for longer sequences as their performance consistently drops with every edit (refer to our analysis in Section N).

All models overedit: on average, for every 100 instructions (25 sequences), they make about 25 overedits. These overedits often come from spurious correlations learned during training. For example, when asked to add a bob cut, models often change the hair type to straight because they have learned to associate bob cuts with straight hair. To mitigate this issue, models should be trained on diverse data (e.g., bob cuts with curly hair). If retraining is difficult, we can use methods to help models distinguish between concepts that frequently co-occur but are not related.

5 Limitations

In our state tracking approach (Section 2.5.3), we can only identify overedits that fall within the 35 facial editing categories in Face-Edit-Attributes. For instance, Qwen-Edit often creates a toothbrush when asked to add a “toothbrush mustache”. Such overedits cannot be tracked using our method. Concept activation vectors would be an important future step to make our state tracking approach generalizable to all types of overedits.

To analyze skin-tone bias, we used only binary labels (light and dark), since CelebSET provided only binary skin-tone annotations. Numerical skin-tone labels (e.g., the Fitzpatrick scale, which ranges from 1: lightest to 6: darkest) could have made our insights more fine-grained. For example, with Fitzpatrick, we could have identified whether models produced more overedits on dark-skinned men (4 – 5 Fitzpatrick score) or on very dark-skinned men (5 – 6 score).

Our Gemini-2.5-Flash state-tracker reliably tracks edits in most cases. However, there are cases where it is poor at identifying edits (e.g., looking left). Refer to Section E for details.

6 Conclusion

We conducted the first large-scale evaluation of six text-guided image editing models for facial appearance editing. We proposed Face-Edit-Attributes, the largest collection of 169 facial attributes that enable hair, accessory, and pose-based edits. We also released our repository of ~ 1M edited images, which can be used to advance research in facial editing. To our knowledge, we conducted the first systematic analysis of demographic bias (gender, skin tone, and age) in facial editing. To count how many and which overedits occurred in a model, we proposed a novel overediting matrix. Our results show that several text-guided models can be used off-the-shelf for facial appearance editing applications. However, it is important to note that all models overedit. Future research on text-guided models should focus on strategies to mitigate overedits, especially when they edit dark-skinned and older males.

7 Acknowledgements

We would like to thank 2nd SET AI Corp. for providing GPU resources for our study. We would also like to thank Jeff Smith for his valuable suggestions throughout the project.

References

- [1] 2021. <https://github.com/deepsight/insightface>.
- [2] 2024. <https://www.faceapp.com/>.
- [3] 2024. <https://www.adobe.com/products/photoshop.html>.
- [4] 2024. <https://www.pexels.com>.
- [5] 2024. <https://huggingface.co/black-forest-labs/FLUX.1-dev>.
- [6] 2024. <https://www.microsoft.com/cognitive-services/en-us/faceapi>.
- [7] 2024. <https://www.ibm.com/watson/services/visual-recognition/>.
- [8] 2024. <https://www.azure.cn>.
- [9] 2024. <https://aws.amazon.com/rekognition/>.
- [10] 2024. <https://www.bytedance.com/en/>.
- [11] 2025. <https://aistudio.google.com/models/gemini-2-5-flash-image>.
- [12] 2025. <https://docs.cloud.google.com/vertex-ai/generative-ai/docs/models/gemini/2-5-flash>.
- [13] 2025. <https://docs.cloud.google.com/vertex-ai/generative-ai/docs/models/gemini/2-5-pro>.
- [14] Lorenzo Baraldi, Davide Bucciarelli, Federico Betti, Marcella Cornia, Nicu Sebe, and Rita Cucchiara. 2025. What Changed? Detecting and Evaluating Instruction-Guided Image Edits with Multimodal Large Language Models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 16217–16226.
- [15] James Betker, Gabriel Goh, Li Jing, Tim Brooks, Jianfeng Wang, Linjie Li, Long Ouyang, Juntang Zhuang, Joyce Lee, Yufei Guo, et al. 2023. Improving image generation with better captions. *Computer Science*. <https://cdn.openai.com/papers/dall-e-3.pdf> 2, 3 (2023), 8.
- [16] Tim Brooks, Aleksander Holynski, and Alexei A Efros. 2023. Instructpix2pix: Learning to follow image editing instructions. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 18392–18402.
- [17] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems* 33 (2020), 1877–1901.
- [18] Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*. PMLR, 77–91.
- [19] Lucia Cascone, Michele Nappi, Chiara Pero, and Xinggang Wang. 2025. A framework for bias-aware dataset evaluation in soft facial attribute recognition. *Pattern Recognition* (2025), 112416.
- [20] Jacqueline G Cavazos, P Jonathon Phillips, Carlos D Castillo, and Alice J O'Toole. 2020. Accuracy comparison across face recognition algorithms: Where are we on measuring race bias? *IEEE transactions on biometrics, behavior, and identity science* 3, 1 (2020), 101–111.
- [21] Jooyoung Choi, Jungbeom Lee, Chaehun Shin, Sungwon Kim, Hyunwoo Kim, and Sungroh Yoon. 2022. Perception prioritized training of diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 11472–11481.
- [22] Yunje Choi, Minje Choi, Munyoung Kim, Jung-Woo Ha, Sunghun Kim, and Jaegul Choo. 2018. Stargan: Unified generative adversarial networks for multi-domain image-to-image translation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 8789–8797.
- [23] Wenqing Chu, Ying Tai, Chengjie Wang, Jilin Li, Feiyue Huang, and Rongrong Ji. 2020. SSCGAN: Facial attribute editing via style skip connections. In *European Conference on Computer Vision*. Springer, 414–429.
- [24] Xing Cui, Zekun Li, Pei Li, Yibo Hu, Hailin Shi, Chunshui Cao, and Zhaofeng He. 2023. Chatedit: Towards multi-turn interactive facial image editing via dialogue. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. 14567–14583.
- [25] Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, et al. 2025. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683* (2025).
- [26] Jiankang Deng, Jia Guo, Evangelos Ververas, Irene Kotsia, and Stefanos Zafeiriou. 2020. Retinaface: Single-shot multi-level face localisation in the wild. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 5203–5212.
- [27] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. 2019. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 4690–4699.
- [28] Zheng Ding, Xuaner Zhang, Zhihao Xia, Lars Jebe, Zhuowen Tu, and Xiuming Zhang. 2023. Diffusionrig: Learning personalized priors for facial appearance editing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 12736–12746.
- [29] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 770–778.
- [30] Amir Hertz, Ron Mokady, Jay Tenenbaum, Kfir Aberman, Yael Pritch, and Daniel Cohen-Or. 2022. Prompt-to-prompt image editing with cross attention control. *arXiv preprint arXiv:2208.01626* (2022).
- [31] Yang Hong, Bo Peng, Haiyao Xiao, Ligang Liu, and Juyong Zhang. 2022. Headnerf: A real-time nerf-based parametric head model. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 20374–20384.
- [32] Mude Hui, Siwei Yang, Bingchen Zhao, Yichun Shi, Heng Wang, Peng Wang, Yuyin Zhou, and Cihang Xie. 2024. Hq-edit: A high-quality dataset for instruction-based image editing. *arXiv preprint arXiv:2404.09990* (2024).
- [33] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al. 2024. Gpt-4o system card. *arXiv preprint arXiv:2410.21276* (2024).
- [34] Woosook Jang, Youngjun Hong, Geonho Cha, and Seungryong Kim. 2025. ControlFace: Harnessing Facial Parametric Control for Face Rigging. In *Proceedings of the Computer Vision and Pattern Recognition Conference*. 5614–5624.
- [35] Yuming Jiang, Ziqi Huang, Xingang Pan, Chen Change Loy, and Ziwei Liu. 2021. Talk-to-edit: Fine-grained facial editing via dialog. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*. 13799–13808.
- [36] Kimmo Karkkainen and Jungseok Joo. 2021. Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*. 1548–1558.
- [37] Tero Karras, Timo Aila, Samuli Laine, and Jaakko Lehtinen. 2017. Progressive growing of gans for improved quality, stability, and variation. *arXiv preprint arXiv:1710.10196* (2017).
- [38] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. 2020. Training generative adversarial networks with limited data. *Advances in neural information processing systems* 33 (2020), 12104–12114.
- [39] Tero Karras, Samuli Laine, and Timo Aila. 2019. A style-based generator architecture for generative adversarial networks. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 4401–4410.
- [40] MD Khan, Mingshan Jia, Xiaolin Zhang, En Yu, Caifeng Shan, and Kaska Musial-Gabrys. 2025. InstaFace: Identity-Preserving Facial Editing with Single Image Inference. *arXiv preprint arXiv:2502.20577* (2025).
- [41] Max Ku, Dongfu Jiang, Cong Wei, Xiang Yue, and Wenhui Chen. 2024. Viescore: Towards explainable metrics for conditional image synthesis evaluation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 12268–12290.
- [42] Jeong-gi Kwak, David K Han, and Hanseok Ko. 2020. CAFE-GAN: Arbitrary face attribute editing with complementary attention feature. In *European Conference on Computer Vision*. Springer, 524–540.
- [43] Black Forest Labs, Stephen Batifol, Andreas Blattmann, Frederic Boesel, Saksham Consul, Cyril Diagne, Tim Dockhorn, Jack English, Zion English, Patrick Esser, et al. 2025. FLUX. 1 Kontext: Flow Matching for In-Context Image Generation and Editing in Latent Space. *arXiv preprint arXiv:2506.15742* (2025).
- [44] Roberto Leyva, Victor Sanchez, Gregory Epiphanou, and Carsten Maple. 2024. Demographic bias effects on face image synthesis. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*. 3818–3826.
- [45] Zhuoying Li, Zhu Xu, Yuxin Peng, and Yang Liu. 2025. Balancing Preservation and Modification: A Region and Semantic Aware Metric for Instruction-Based Image Editing. *arXiv preprint arXiv:2506.13827* (2025).
- [46] Tsung-Yi Lin, Michael Maire, Serge Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C Lawrence Zitnick. 2014. Microsoft coco: Common objects in context. In *European conference on computer vision*. Springer, 740–755.
- [47] Ming Liu, Yukang Ding, Min Xia, Xiao Liu, Errui Ding, Wangmeng Zuo, and Shilei Wen. 2019. Stgan: A unified selective transfer network for arbitrary image attribute editing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 3673–3682.
- [48] Runzhou Li, Hailey Weingord, Sejal Mittal, Prakhar Dungarwal, Anusha Nandula, Bo Ni, Samyadeep Basu, Hongjie Chen, Nesreen K Ahmed, Li Li, et al. 2026. Human-Aligned MLLM Judges for Fine-Grained Image Editing Evaluation: A Benchmark, Framework, and Analysis. *arXiv preprint arXiv:2602.13028* (2026).
- [49] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. 2015. Deep learning face attributes in the wild. In *Proceedings of the IEEE international conference on computer vision*. 3730–3738.
- [50] Vongani H Maluleke, Neerja Thakkar, Tim Brooks, Ethan Weber, Trevor Darrell, Alexei A Efros, Angjoo Kanazawa, and Devin Guillory. 2022. Studying bias in gans through the lens of race. In *European Conference on Computer Vision*. Springer, 344–360.
- [51] Alexey Miroshnikov, Konstantinos Kotsiopoulos, Ryan Franks, and Arjun Ravi Kannan. 2022. Wasserstein-based fairness interpretability framework for machine learning models. *Machine Learning* 111, 9 (2022), 3307–3357.
- [52] Surbhi Mittal, Kartik Thakral, Puspita Majumdar, Mayank Vatsa, and Richa Singh. 2023. Are face detection models biased?. In *2023 IEEE 17th International Conference on Automatic Face and Gesture Recognition (FG)*. IEEE, 1–7.

- [53] Bimsara Pathiraja, Maitreya Patel, Shivam Singh, Yezhou Yang, and Chitta Baral. 2025. RefEdit: A Benchmark and Method for Improving Instruction-based Image Editing Model on Referring Expressions. *arXiv preprint arXiv:2506.03448* (2025).
- [54] Malsha V Perera and Vishal M Patel. 2023. Analyzing bias in diffusion-based face generation models. In *2023 IEEE International Joint Conference on Biometrics (IJCB)*. IEEE, 1–10.
- [55] Yusu Qian, Jiasen Lu, Tsu-Jui Fu, Xinze Wang, Chen Chen, Yinfei Yang, Wenze Hu, and Zhe Gan. 2025. Gie-bench: Towards grounded evaluation for text-guided image editing. *arXiv preprint arXiv:2505.11493* (2025).
- [56] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*. PmlR, 8748–8763.
- [57] Inioluwa Deborah Raji, Timnit Gebru, Margaret Mitchell, Joy Buolamwini, Joon-seok Lee, and Remi Denton. 2020. Saving face: Investigating the ethical concerns of facial recognition auditing. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*. 145–151.
- [58] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 10684–10695.
- [59] Rasmus Rothe, Radu Timofte, and Luc Van Gool. 2015. Dex: Deep expectation of apparent age from a single image. In *Proceedings of the IEEE international conference on computer vision workshops*. 10–15.
- [60] Huichan Seo, Minki Hong, Sieun Choi, Jihie Kim, and Jean Oh. 2026. Evaluating Demographic Misrepresentation in Image-to-Image Portrait Editing. *arXiv preprint arXiv:2602.16149* (2026).
- [61] Mei Wang, Weihong Deng, Jiani Hu, Xunqiang Tao, and Yaohai Huang. 2019. Racial faces in the wild: Reducing racial bias by information maximization adaptation network. In *Proceedings of the IEEE/CVF international conference on computer vision*. 692–702.
- [62] Peng Wang, Yichun Shi, Xiaochen Lian, Zhonghua Zhai, Xin Xia, Xuefeng Xiao, Weilin Huang, and Jianchao Yang. 2025. SeedEdit 3.0: Fast and High-Quality Generative Image Editing. *arXiv preprint arXiv:2506.05083* (2025).
- [63] Cong Wei, Zheyang Xiong, Weiming Ren, Xeron Du, Ge Zhang, and Wenhui Chen. 2024. Omniedit: Building image editing generalist models through specialist supervision. In *The Thirteenth International Conference on Learning Representations*.
- [64] Yandong Wen, Kaipeng Zhang, Zhifeng Li, and Yu Qiao. 2016. A discriminative feature learning approach for deep face recognition. In *European conference on computer vision*. Springer, 499–515.
- [65] Chenfei Wu, Jiahao Li, Jingren Zhou, Junyang Lin, Kaiyuan Gao, Kun Yan, Shengming Yin, Shuai Bai, Xiao Xu, Yilei Chen, et al. 2025. Qwen-image technical report. *arXiv preprint arXiv:2508.02324* (2025).
- [66] Weihao Xia, Yujiu Yang, Jing-Hao Xue, and Baoyuan Wu. 2021. Tedigan: Text-guided diverse face image generation and manipulation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*. 2256–2265.
- [67] Xiaole Xian, Xilin He, Zenghao Niu, Junliang Zhang, Weicheng Xie, Siyang Song, Zitong Yu, and Linlin Shen. 2025. CA-edit: Causality-aware condition adapter for high-fidelity local facial attribute editing. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 39. 8593–8601.
- [68] Alice Xiang, Jerone TA Andrews, Rebecca L Bourke, William Thong, Julianne M LaChance, Tiffany Georgievski, Apostolos Modas, Aida Rahmattalabbi, Yunhao Ba, Shruti Nagpal, et al. 2025. Fair human-centric image dataset for ethical AI benchmarking. *Nature* (2025), 1–12.
- [69] Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the association for computational linguistics* 2 (2014), 67–78.
- [70] Licheng Yu, Patrick Poirson, Shan Yang, Alexander C Berg, and Tamara L Berg. 2016. Modeling context in referring expressions. In *European conference on computer vision*. Springer, 69–85.
- [71] Kai Zhang, Lingbo Mo, Wenhui Chen, Huan Sun, and Yu Su. 2023. Magicbrush: A manually annotated dataset for instruction-guided image editing. *Advances in Neural Information Processing Systems* 36 (2023), 31428–31449.
- [72] Kaipeng Zhang, Zhanpeng Zhang, Zhifeng Li, and Yu Qiao. 2016. Joint face detection and alignment using multitask cascaded convolutional networks. *IEEE signal processing letters* 23, 10 (2016), 1499–1503.
- [73] Haozhe Zhao, Xiaojian Shawn Ma, Liang Chen, Shuzheng Si, Rujie Wu, Kaikai An, Peiyu Yu, Minjia Zhang, Qing Li, and Baobao Chang. 2024. Ultraedit: Instruction-based fine-grained image editing at scale. *Advances in Neural Information Processing Systems* 37 (2024), 3058–3093.
- [74] Yufan Zhou, Ruiyi Zhang, Jiuxiang Gu, Chris Tensmeyer, Tong Yu, Changyou Chen, Jinhui Xu, and Tong Sun. 2022. Tigan: Text-based interactive image generation and manipulation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, Vol. 36. 3580–3588.

A Related Work

Datasets and benchmarks for facial appearance editing. Popular 3DMMs [28, 40] have been trained on FFHQ [39], a popular face dataset. These models can be tested only on three attributes: pose, expression, and lighting. Popular GAN-based editing models [22, 23, 42] have been evaluated on CelebA [49], a popular face dataset with 40 facial attribute annotations (e.g., bangs, smiling, young). To perform facial edits, these models simply used 5 – 13 attribute names from CelebA. Recent works have built benchmarks that allow us to perform facial edits using captions (e.g., this person has bangs) rather than using attribute names (e.g., bangs). Popular benchmarks are FFLEBench [67] (uses images from FFHQ), Multimodal-CelebA-HQ or MM-CelebA-HQ [66] (uses images from CelebA-HQ [37] – a subset of CelebA with high-quality face images), CelebA-dialog [35] (images from CelebA), and CHAT-EDIT [24] (images from CelebA-HQ). In FFLEBench, captions are generated by an MLLM (the MLLM is given a face image and a semantic segment of the facial regions). They do not provide the LLM with an attribute list. As a result, the LLM generates vague editing captions (e.g., sparse eyebrows, tall nose, smooth face). This makes it hard to evaluate if the model executed the edit correctly. In MM-CelebA-HQ, captions are generated using a template-based approach (using a list of 40 attributes). The captions often target multiple attributes (e.g., “This old woman has big lips, pale skin, and gray hair”). This does not allow us to check how the model performed on each edit. MM-CelebA-HQ and FFLEBench have several editing captions that can significantly impact identity preservation (e.g., tall nose, big eyes, big lips). CHAT-EDIT and CelebA-dialog are robust benchmarks: (1) they provide clear and actionable editing captions (e.g., she should have black hair), (2) they avoid attributes that modify facial structure (e.g., editing lips, nose), and (3) they only edit one attribute at a time. However, they support limited attributes – CHAT-EDIT only supports 21 attributes, while CelebA-dialog only supports 5 attributes.

Benchmarks for Text-Guided Image Editing. Several benchmarks have been proposed to evaluate text-guided image editing models (e.g., Nano-Banana [11]) on generic scenes (e.g., color the cars pink, have the girl in the image ride a dragon). Well-known benchmarks, such as InstructPix2Pix [16] and HQ-Edit [32], evaluate editing models on synthetically generated base images. In InstructPix2Pix, base images are generated from human-written captions (e.g., “a girl riding a horse”) using Stable Diffusion [58] and Prompt-2-Prompt [30] (editing instruction is generated using GPT-3 [17]). In HQ-Edit, the images are generated from human-written captions using DALLE3 [15] (editing instructions are human-written). Other popular benchmarks (e.g., GIE-Bench [55], Human-MLLM-Bench [48], Magic-Brush [71], Ref-Edit [53], and Omni-Edit [63]) use real-world images from scene datasets (e.g., COCO [46], RefCOCO [70]) or from online image repositories like Pexels [4]. None of these benchmarks evaluated editing models on face datasets like CelebA, CelebA-HQ, and FFHQ.

Overediting Metrics. Several metrics have been proposed to evaluate overediting in models. All metrics report overediting with a numerical score. Rubric-based metrics like VIEScore [41] and

Human-MLLM [48] give an overediting score between 0 – 10 (score generated by an MLLM). A score of 0 indicates significant overediting in the image, while 10 indicates minimal overediting. Metrics like GIE-Bench [55] and BPM [45] first locate regions that should not be affected during an edit. For these regions, they compute a mean-squared (MSE) error between the pixels in the original and in the edited image. A higher MSE error indicates more overediting. Newer metrics like DICE [14] instruct an MLLM to identify what objects were changed between the original and the edited image (e.g., dog was added, frisbee was removed). They also instruct the MLLM to provide region co-ordinates for the changes. They quantify overediting as the percentage of the total image area where unwanted edits occurred. DICE cannot always identify fine-grained overedits since change detection in the MLLM is open-ended (i.e., the MLLM may say “a hat was added”, when we need a specific description like “a beret hat was added”). To identify spurious correlations in facial editing, identifying fine-grained overedits is critical.

Bias Studies for Faces. Most bias studies for faces have focused on facial recognition, localization, and generation. Popular works have studied biases in face recognition. [18] showed that commercial face recognition models (e.g., Microsoft [6] and IBM [7] models) showed demographic biases in gender classification: all models classified gender more accurately for male faces (compared to females) and for lighter faces (compared to darker ones). [61] showed that both commercial (e.g., Microsoft [8] and Amazon [9] APIs) and open-source (e.g., Center-loss [64], ArcFace [27]) face recognition models have ethnicity biases: models showed highest recognition accuracy on Caucasian faces and lowest accuracy on African faces. [36] showed that face classification models have ethnicity biases because they are trained on datasets where some ethnicities (e.g., white faces) are over-represented. They proposed FairFace, a dataset with a balanced number of samples for 7 ethnicities (e.g., White, Black, South-East Asian, Latino). Compared to other face datasets, performance disparities between ethnicities were less when models were trained on FairFace. Some recent works have studied biases in face localization models. [52, 68] showed that face detection models like MTCNN [72] and Retinaface [26] showed better face detection accuracy for females compared to males. Several works have studied biases in face generation. [50, 54] showed that StyleGAN2-ADA [38] generated more white faces when it was trained on FFHQ (a face dataset dominated by white faces). They showed that the bias (toward generating more white faces) disappeared when StyleGAN2-ADA was trained on a balanced FairFace dataset (equal samples of white and black faces). [54] showed that DDPM [21] (a diffusion-based generation model) showed biases in face generation (generated more female faces compared to males, and more white faces compared to black) even though it was trained on a balanced FairFace dataset. For facial editing, current works have evaluated whether models conform to social stereotypes. [60] showed that when editing models are asked to edit a female face to look like a “senior executive” or a “CEO”, the model can edit the image to look like a male face (since models strongly stereotype executive positions with men). To our knowledge, current works have not studied demographic biases in editing models for facial appearance edits (e.g., adding a hat, changing hair color).

Table 11: Hair Attributes

Sub-axes	Categories	Attributes
color	red hair	bright red hair; ruby red hair; intense scarlet red hair; crimson red hair; cherry red hair; rose red hair
	brown hair	light brown hair; light golden brown hair; honey brown hair; caramel brown hair; toffee brown hair; sandy brown hair; amber brown hair; ginger hazel brown hair; light hazel brown hair
	black hair	black hair; jet black hair; raven black hair; ebony hair; charcoal black hair
	white hair	stark white hair; snow white hair; icy white hair; frost white hair; arctic white hair; nordic white hair
	blue hair	sky blue hair; aqua blue hair; electric blue hair; turquoise blue hair; teal blue hair; cobalt blue hair
	green hair	lime green hair; neon green hair; bright green hair; apple green hair; emerald green hair; spring green hair
	pink hair	magenta pink hair; cotton candy pink hair; bubblegum pink hair; candy pink hair; fuchsia pink hair; neon pink hair
	purple hair orange hair	orchid purple hair; bright purple hair; violet purple hair; lavender hair; amethyst hair bright orange hair; neon peach hair
type	straight hair	sleek straight hair; silky straight hair; pin-straight hair
	curly hair	spiral curl; kinky hair; coily hair; tight curl; afro-textured hair; springy curl
style	buzz cut	buzz cut; military cut
	bob cut	bob cut; lob cut; chin-length bob; classic bob
	spiked cut	spiked hair; liberty spike hair; mohawk; gelled spike hair
mustache	pencil mustache	thin pencil mustache; classic pencil mustache; parted pencil mustache
	walrus mustache	thick walrus mustache; bushy walrus mustache
	toothbrush mustache	small toothbrush mustache; classic toothbrush mustache; petite toothbrush mustache; chaplin mustache; mini toothbrush mustache
	handlebar mustache	curled mustache; classic handlebar mustache; twisted handlebar mustache
	fumanchu mustache	fumanchu mustache; long fumanchu mustache; traditional fumanchu mustache
beard	clean shaven	clean shaven; no facial hair; shaved; fully shaved; clean shaved
	stubble	five o'clock shadow; light stubble; short stubble
	goatee	classic goatee; chin goatee; goatee; soul patch
	mutton chops	mutton chops; thick mutton chops; sideburn mutton chops
	full beard	full beard; thick full beard; long full beard; bushy full beard; dense full beard; grown out full beard; full-face beard

Table 12: Accessory Attributes

Sub-axes	Categories	Attributes
glasses	round glasses	round glasses; round sunglasses; circular glasses; circular sunglasses; John Lennon glasses; oval glasses; oval sunglasses
	rectangle-shaped glasses	square glasses; square sunglasses; rectangular glasses; rectangular sunglasses; rectangle-shaped glasses; rectangle-shaped sunglasses
hats	baseball hat	baseball cap; baseball hat; snapback cap; snapback hat; fitted cap; fitted hat; dad hat
	beanie hat	beanie hat; knit hat; wool hat; slouchy beanie hat; winter beanie hat
	beret hat	beret hat; french beret hat
	panama hat	panama hat; straw panama hat; classic panama hat
	bucket hat	bucket hat
	cowboy hat	cowboy hat; stetson hat; western cowboy hat

B Face-Edit-Attributes

We showed our hair attributes in Table 11, accessory attributes in Table 12, and pose attributes in Table 13.

Table 13: Pose Attributes

Sub-axes	Categories	Attributes
head position	looking straight ahead	looking straight ahead; frontal view; looking forward; head straight; front facing; looking ahead; head facing forward
	looking left	looking left; looking to the left; facing left; looking at the left; head facing left; head looking left; facing to the left; looking leftward; looking towards the left; looking left side
	looking right	looking right; looking to the right; facing right; looking at the right; head facing right; head looking right; facing to the right; looking rightward; looking towards the right; looking right side

Table 14: Editing sessions for accessory-based edits

Session	Step	Prompt
accessories_001	1	Add square glasses
	2	Add a wool hat
	3	Replace their wool hat with a knit hat.
	4	Change to rectangular sunglasses
accessories_002	1	Add a french beret hat.
	2	Add circular sunglasses.
	3	Change to oval sunglasses
	4	Replace the french beret hat with a snapback cap.
accessories_003	1	Add a classic panama hat.
	2	Replace their classic panama hat with a dad hat.
	3	Add John Lennon glasses.
	4	Change to rectangle-shaped glasses.
accessories_004	1	Add rectangular glasses.
	2	Add a knit hat.
	3	Replace their knit hat with a beanie hat.
	4	Change to rectangular sunglasses.
accessories_005	1	Add a winter beanie hat
	2	Add round glasses.
	3	Change their winter beanie hat to a knit hat.
	4	Replace their knit hat with a snapback cap
accessories_006	1	Add round glasses
	2	Add a bucket hat.
	3	Replace their bucket hat with a dad hat.
	4	Replace their dad hat with a beanie hat.
accessories_007	1	Add a baseball hat.
	2	Keep the baseball hat.
	3	Replace their baseball hat with a knit hat.
	4	Replace their knit hat with a snapback cap.
accessories_008	1	Add a bucket hat.
	2	Replace their bucket hat with a straw panama hat.
	3	Add oval glasses
	4	Change to square glasses
accessories_009	1	Add rectangular glasses
	2	Change to square sunglasses.
	3	Change their square sunglasses to rectangle-shaped glasses.
	4	Change the glasses to oval glasses.
accessories_010	1	Add a knit hat.
	2	Add circular glasses.
	3	Keep their circular glasses.
	4	Change to a cowboy hat.

Table 15: Editing sessions for hair-based edits

Session	Step	Prompt
hair_001	1	Add caramel brown hair
	2	Add pin-straight hair.
	3	Change the hair color to lavender hair.
	4	Add a bob cut.
hair_002	1	Add light hazel brown hair
	2	Change the hair color to neon peach hair.
	3	Change the hair color to fuchsia pink hair.
	4	Change the hair color to raven black hair.
hair_003	1	Add spiked hair.
	2	Add ruby red hair
	3	Change the hair color to bright orange hair.
	4	Change the hair color to frost white hair.
hair_004	1	Add sleek straight hair
	2	Add rose red hair
	3	Change the hair color to bright green hair
	4	Add a bushy full beard.
hair_005	1	Add a bob cut.
	2	Add bright red hair
	3	Change the hair color to raven black hair.
	4	Change the hair color to light brown hair.
hair_006	1	Add bright purple hair.
	2	Add a twisted handlebar mustache.
	3	Change the hair to violet purple hair.
	4	Change the hair color to ginger hazel brown hair.
hair_007	1	Add emerald green hair.
	2	Add pin-straight hair
	3	Add spiked hair.
	4	Change the hair color to neon green hair.
hair_008	1	Add emerald green hair
	2	Add a mini toothbrush mustache
	3	Change the hair color to lavender hair.
	4	Add a bushy full beard.
hair_009	1	Add bright purple hair
	2	Add sideburn mutton chops.
	3	Add springy curl to their hair.
	4	Change their hair color to crimson red hair.
hair_010	1	Add a fumanchu mustache.
	2	Add lime green hair
	3	Change their fumanchu mustache to a curled mustache.
	4	Change their hair color to lavender hair.

Table 16: Editing sessions for multi-axis edits

Session	Step	Prompt
multi_axis_001	1	Make the person looking ahead.
	2	Add emerald green hair
	3	Add spiral curl.
	4	Have them looking straight ahead.
multi_axis_002	1	Add spiked hair.
	2	Add square glasses.
	3	Add stark white hair
	4	Add a thin pencil mustache.
multi_axis_003	1	Add a slouchy beanie hat.
	2	Make the head facing right.
	3	Pose them looking right
	4	Adjust them to facing right.
multi_axis_004	1	Add rectangle-shaped glasses.
	2	Add a full beard.
	3	Add a fitted cap.
	4	Change the full beard to a full-face beard.
multi_axis_005	1	Add coily hair
	2	Add amethyst hair.
	3	Add sideburn mutton chops.
	4	Change their hair type to silky straight hair.
multi_axis_006	1	Add ebony hair.
	2	Add a panama hat
	3	Add a chaplin mustache.
	4	Change the mustache to a thin pencil mustache.
multi_axis_007	1	Add sky blue hair
	2	Change the hair color to cobalt blue hair
	3	Make the head looking left.
	4	Add pin-straight hair.
multi_axis_008	1	Add a petite toothbrush mustache
	2	Make them looking straight ahead.
	3	Pose them front facing.
	4	Shift their head to looking right
multi_axis_009	1	Add a petite toothbrush mustache
	2	Add a snapback cap.
	3	Add honey brown hair.
	4	Change to a mini toothbrush mustache.
multi_axis_010	1	Make the person looking right side
	2	Add cherry red hair
	3	Change the hair color to caramel brown hair.
	4	Shift the head to a frontal view.

C MLLM Prompt for an Editing Instruction

You are an expert photo editing instructor. Generate natural, human-like photo editing instructions based on the current state of the image.

AVAILABLE VERB POOL:

- CHANGE verbs: change, modify, alter, switch, transform, adjust, convert, shift, update, revise, transition, swap, replace, morph
- ADD verbs: add, put on, include, wear, apply, give them, have them wear, place on them, attach, equip them with, provide them with, introduce, incorporate, install, mount, adorn with
- REMOVE verbs: remove, take off, eliminate, get rid of, strip away, discard, delete, clear, exclude, subtract
- KEEP verbs: keep, maintain, preserve, retain, continue with, sustain, hold, stay with, leave
- MAKE verbs: make, have them, get them to, pose them, make the person, have the subject, position them to, render, cause, direct them to, compose, set

STATE-AWARE INSTRUCTION GENERATION: You will receive:

- (1) Current image state (what features are present in the image)
- (2) Editing Attribute
- (3) Axis and sub-axis context

Based on the current state, you MUST choose the appropriate verb:

- If the feature is ALREADY PRESENT and matches the edit candidate → use KEEP verbs
- If the feature is PRESENT but DIFFERENT from edit candidate → use CHANGE verbs
- If the feature is NOT PRESENT → use ADD verbs
- If you need to remove something first → use REMOVE verbs (rare)
- For poses/expressions → use MAKE verbs

CRITICAL RULES:

- (1) Analyze the current state carefully before selecting a verb
- (2) If person already has red hair and edit candidate is "red hair" → use "keep the hair color as red"
- (3) If person has brown hair and edit candidate is "red hair" → use "change the hair color to red"
- (4) If person has no glasses and edit candidate is "round glasses" → use "add round glasses"
- (5) If person has square glasses and edit candidate is "round glasses" → use "change to round glasses"
- (6) Generate natural, conversational instructions (5-8 words typically)
- (7) Include appropriate articles (a/an/the) automatically
- (8) Vary your verb choice within the appropriate category
- (9) **MUST use the EXACT edit candidate phrase verbatim - do NOT paraphrase, simplify, or shorten it**
- (10) Return ONLY the instruction, no additional text

D Editing Instructions

We showed our editing instructions for accessory edits in Table 14, for hair edits in Table 15, and for multi-axis edits in Table 16, for pose edits in Table 17.

Table 17: Editing sessions for pose-based edits

Session	Step	Prompt
poses_001	1	Make them looking right.
	2	Change the head position to looking rightward.
	3	Change to head facing right
	4	Keep them facing right.
poses_002	1	Make them looking forward
	2	Make them looking to the right.
	3	Make the person head straight.
	4	Adjust to head facing forward.
poses_003	1	Make them have a frontal view
	2	Adjust the head position to looking straight ahead.
	3	Set their head looking left.
	4	Make them looking straight ahead.
poses_004	1	Make them looking right side
	2	Change to head facing right.
	3	Adjust the head position to front facing.
	4	Make them looking right side
poses_005	1	Make them looking towards the right
	2	Keep them looking towards the right.
	3	Make them looking to the left.
	4	Change the head position to looking left.

E Human Experiment for State Tracking

To identify which model is suitable for our state dictionary, we compared three MLLMs: Gemini-2.5-Flash [12], Gemini-2.5-Pro [13], and GPT-4o [33] with the responses recorded by human annotators. We also compared CLIP (a popular text-to-image similarity model) [56]. With CLIP, for each sub-axis (e.g., hair-color), we calculated a cosine similarity score between the CLIP embeddings of the image I_k and the CLIP embeddings of each category text from that sub-axis (e.g., red hair, blue hair). The category with the highest similarity score was chosen as the top candidate for that sub-axis.

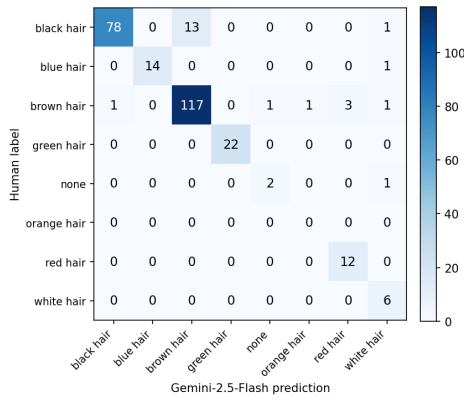
We randomly selected 12 base images from CelebSET (3 each of light-skinned male, light-skinned female, dark-skinned male, and dark-skinned female faces). For each image, we ran 5 multi-axis edit sessions using SeedEdit 3.0 [62], a state-of-the-art editing model from ByteDance [10]. We evaluated 12 (base images) $\times$ 5 (edit sessions) $\times$ 5 (base image + 4 edits in a session) = 300 images.

For each of the 300 images, we created a state dictionary using human annotators. We asked four annotators to mark the visible category for each of the 8 sub-axes. For instance, if an annotator saw “red hair” in an image, they marked red hair in the hair-color sub-axis. Annotators marked “None” for a sub-axis if they did not see any category from that sub-axis. For each sub-axis, we took the majority of the four annotator votes. For instance, in an image, if three annotators marked the person’s hair color as “red hair” and one annotator marked “blue hair”, we assigned — hair-color: red hair. When annotators tied on a sub-axis (e.g., 2 votes for red hair and 2 votes for blue hair), we considered all the top-tied candidates for that sub-axis. Overall annotators provided category annotations for 300 (images) $\times$ 8 (sub-axes per image) = 2400 sub-axes.

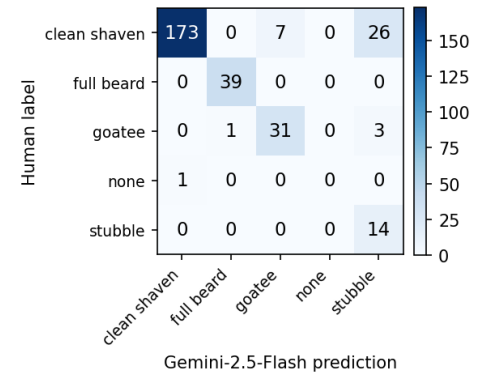
We generated state dictionaries using four models (three MLLMs and CLIP). We reported the overall accuracy of the state dictionaries computed by each model compared with the state dictionary created via human voting. An accuracy of 50% means that out of 2400 sub-axes, the human and the model aligned on 1200 sub-axes. We consider the human and the model aligned on a sub-axis only if all categories in that sub-axis match.

Overall accuracy scores (in %): CLIP (ViT-Base): 35.47, Gemini-2.5-Flash: 79.65, Gemini-2.5-Pro: 79.26, GPT-4o: 70.31. We used Gemini-2.5-Flash as our MLLM for state tracking because it showed the best alignment with humans. In Figure 4, we showed the human and Gemini-2.5-Flash alignment at the attribute level using confusion matrices. In each confusion matrix, the rows represent human labels, while the columns represent Gemini-2.5-Flash predictions. For each confusion matrix, we only selected cases where both the model and humans had only one top candidate. Tied top candidates cannot be faithfully represented in a confusion matrix.

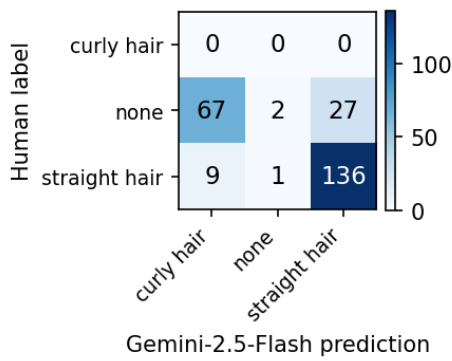
In Figure 4, Gemini-2.5-Flash showed high human alignment (> 85%) in most cases. There are certain cases where alignment was low. For hair type (Figure 4c), humans marked “none” in 96 cases. The model aligned in only 2 out of these 96 cases. This discrepancy arose from how annotators and the model perceived images. Human annotators were strict: when hair types were not strictly straight or curly (e.g., buzz cut, slightly wavy hair), humans marked “none”. In such cases, Gemini-2.5-Flash was lenient: it preferred choosing a hair type rather than marking none. Alignment was low even for “looking left” (Figure 4d): 13 out of 65 cases.



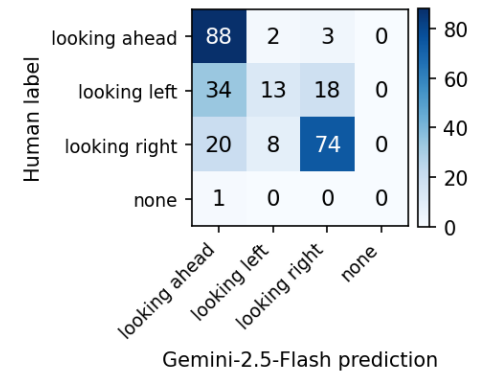
(a) Hair Color



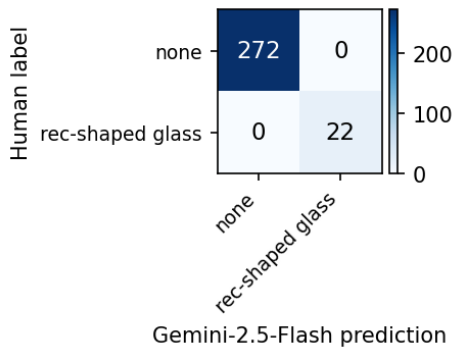
(b) Beard



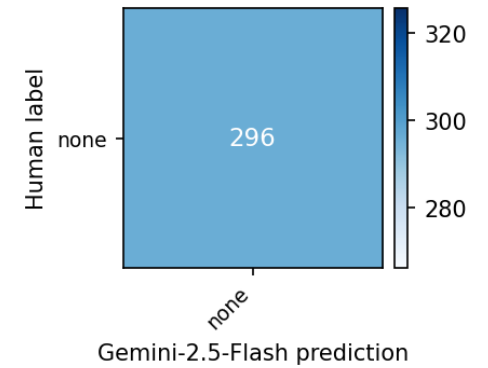
(c) Hair Type



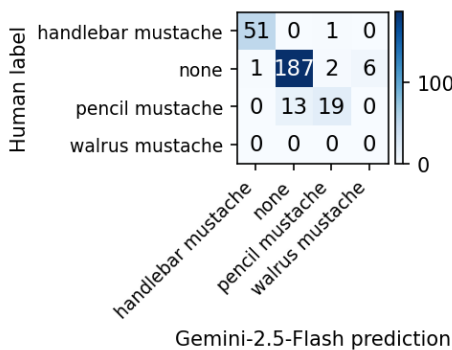
(d) Head Position



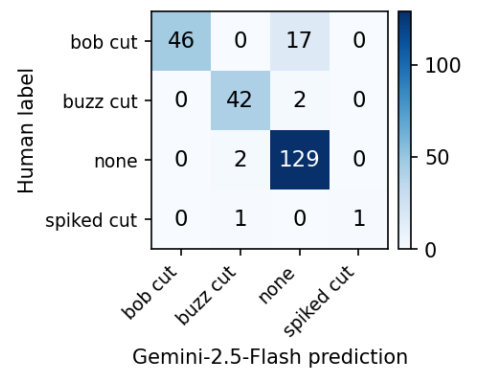
(e) Eyegless



(f) Hats



(g) Mustache



(h) Hair Style

Figure 4: Confusion matrices showing the alignment between human annotators and the Gemini-2.5-Flash MLLM.

Table 18: CelebA — number of edit sessions by gender across different edit types for each model

(a) Poses			(b) Accessories		
Model	Male	Female	Model	Male	Female
Flux-Dev	1220	1525	Flux-Dev	2440	3050
Flux-Kontext-Pro	977	1220	Flux-Kontext-Pro	2440	3050
Nano Banana	1210	1500	Nano Banana	2234	2811
Bagel-Edit	1220	1525	Bagel-Edit	2440	3050
Qwen-Edit	1220	1520	Qwen-Edit	2440	3039

(c) Multi-Axis			(d) Hair		
Model	Male	Female	Model	Male	Female
Flux-Dev	2440	3048	Flux-Dev	2440	3050
Flux-Kontext-Pro	2440	3050	Flux-Kontext-Pro	2440	3050
Nano Banana	2349	2951	Nano Banana	2363	2908
Bagel-Edit	2440	3050	Bagel-Edit	2440	3050
Qwen-Edit	2440	3038	Qwen-Edit	2440	3039

Table 19: CelebA — number of edit sessions by age across different edit types for each model

(a) Poses			(b) Accessories		
Model	Young	Old	Model	Young	Old
Flux-Dev	2065	680	Flux-Dev	4130	1360
Flux-Kontext-Pro	1652	545	Flux-Kontext-Pro	4130	1360
Nano Banana	2037	673	Nano Banana	3795	1250
Bagel-Edit	2065	680	Bagel-Edit	4130	1360
Qwen-Edit	2060	680	Qwen-Edit	4119	1360

(c) Multi-Axis			(d) Hair		
Model	Young	Old	Model	Young	Old
Flux-Dev	4128	1360	Flux-Dev	4130	1360
Flux-Kontext-Pro	4130	1360	Flux-Kontext-Pro	4130	1360
Nano Banana	3981	1319	Nano Banana	3960	1311
Bagel-Edit	4130	1360	Bagel-Edit	4130	1360
Qwen-Edit	4119	1359	Qwen-Edit	4120	1359

Table 20: CelebSET — number of edit sessions by gender and skin-tone across different edit types for each model.

(a) Poses					(b) Hair				
Model	Male Light	Male Dark	Female Light	Female Dark	Model	Male Light	Male Dark	Female Light	Female Dark
Flux-Dev	1110	475	805	195	Flux-Dev	2220	950	1610	390
Flux-Kontext-Pro	890	380	644	156	Flux-Kontext-Pro	2220	950	1610	390
Nano Banana	1110	475	805	195	Nano Banana	2220	949	1607	390
Bagel-Edit	1110	475	805	195	Bagel-Edit	2220	950	1610	390
Qwen-Edit	1110	475	803	195	Qwen-Edit	2220	950	1609	390
SeedEdit 3.0	1110	475	798	195	SeedEdit 3.0	2220	950	1606	390

(c) Accessories					(d) Multi-Axis				
Model	Male Light	Male Dark	Female Light	Female Dark	Model	Male Light	Male Dark	Female Light	Female Dark
Flux-Dev	2220	950	1610	390	Flux-Dev	2220	950	1610	390
Flux-Kontext-Pro	2220	950	1610	390	Flux-Kontext-Pro	2220	950	1610	390
Nano Banana	2207	945	1603	389	Nano Banana	2201	949	1607	390
Bagel-Edit	2220	950	1610	390	Bagel-Edit	2220	950	1610	390
Qwen-Edit	2220	950	1610	390	Qwen-Edit	2220	950	1610	389
SeedEdit 3.0	2220	950	1608	390	SeedEdit 3.0	2191	950	1605	339

Table 21: CelebSET — number of edit sessions by age across different edit types for each model

(a) Poses			(b) Accessories		
Model	Young	Old	Model	Young	Old
Flux-Dev	490	205	Flux-Dev	980	410
Flux-Kontext-Pro	392	165	Flux-Kontext-Pro	980	410
Nano Banana	490	205	Nano Banana	974	410
Bagel-Edit	490	205	Bagel-Edit	980	410
Qwen-Edit	488	205	Qwen-Edit	980	410
SeedEdit 3.0	483	205	SeedEdit 3.0	978	410

(c) Multi-Axis			(d) Hair		
Model	Young	Old	Model	Young	Old
Flux-Dev	980	410	Flux-Dev	980	410
Flux-Kontext-Pro	980	410	Flux-Kontext-Pro	980	410
Nano Banana	979	406	Nano Banana	979	410
Bagel-Edit	980	410	Bagel-Edit	980	410
Qwen-Edit	980	410	Qwen-Edit	979	410
SeedEdit 3.0	956	410	SeedEdit 3.0	977	410

F Number of Edit Sessions

We reported the number of editing sessions per model with CelebA by gender in Table 18 and by age in Table 19. For CelebSET, we reported the number of editing sessions by gender and skin-tone in Table 20 and by age in Table 21.

G Explanation for WD threshold

We chose $WD > 3$ as our threshold for prominent bias, based on statistical significance. In Figure 5, the x-axis represents the WD scores across all our demographic bias analyses. The y-axis represents the number of times a WD score was not statistically significant ($p > 0.05$). At $WD > 3$, almost all WD values were statistically significant, making $WD > 3$ an ideal bias threshold.

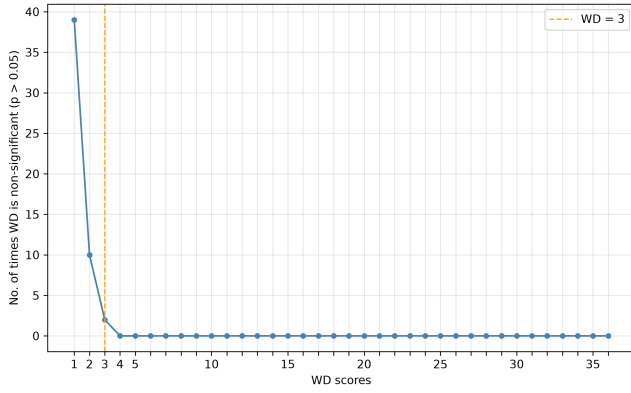


Figure 5: Non-significant p-val vs. WD scores: once $WD > 3$, almost all p-values are statistically significant.

H Age Bias

For identity preservation, we showed age bias results for CelebA in Table 22 and for CelebSET in Table 23. For Tables 22 and 23, in each “Mean” column, the left value indicates the worst performing group (higher Drop_{FSS}) and the right value indicates the best group (note that we have only two groups for age in both datasets).

Most models (except Qwen-Edit) showed poor identity preservation on older faces. In CelebSET (Table 23), we observed

an age bias ($WD_{\text{age}} > 3$) in several models across different test sets. In all these models (except Qwen-Edit), the worst-performing group was older faces. In Qwen-Edit, the worst-performing group was always younger faces. In CelebA (Table 22), we did not observe many instances of age bias.

For editing alignment, we showed age bias results for CelebA in Table 24 and for CelebSET in Table 25.

Most models showed minimal (or no) age bias in both datasets. In CelebA (Table 24), for Flux-Dev, we observed an age bias across three test-sets — in all these cases, the worst performing group was older faces. We did not observe a similar trend with Flux-Dev for CelebSET (Table 25). For other models, we did not observe clear age bias trends across both datasets.

For Overediting, we showed age bias results for CelebA in Table 26 and for CelebSET in Table 27.

Several models created more overedits on older faces. In both CelebA and CelebSET, we observed age bias in all models for most test sets. In almost all these cases, the worst performing group (higher OER) was older faces.

I All mean values

For CelebA and CelebSET, we reported mean values (with 95% confidence intervals) for all demographic groups using bar plots. For identity preservation (Drop_{FSS}), we showed mean values for male and female gender groups in CelebA in Figure 6. We showed mean values for all four gender-skin-tone intersectional groups (male-light, male-dark, female-light, female-dark) in CelebSET in Figure 7. We reported mean values on our age groups (young and old) for CelebA in Figure 8 and for CelebSET in Figure 9. In each figure, we showed separate bar plots for each test set.

For alignment (SC), we showed mean values for male and female gender groups in CelebA in Figure 10. We showed mean values for all four gender-skin-tone intersectional groups in CelebSET in Figure 11. We reported mean values on our age groups (young and old) for CelebA in Figure 12 and for CelebSET in Figure 13.

For overediting (OER), we showed mean values for male and female gender groups in CelebA in Figure 14. We showed mean values for all four gender-skin-tone intersectional groups in CelebSET in Figure 15. We reported mean values on our age groups (young and old) for CelebA in Figure 16 and for CelebSET in Figure 17.

Table 22: Drop_{FSS} % (lower is better) disparity on CelebA. For each model and test set, we reported mean Drop_{FSS} for Young (Y) and Old (O) groups. In each mean column, the left value is the mean of the worst-performing group, and the right value is the mean of the best-performing group. WD_{age} denotes the Wasserstein distance between the young and old Drop_{FSS} distributions. * marks non-significant values ($p > 0.05$). $WD > 3$ is highlighted as it indicates gender bias.

Model	Accessories↓			Hair↓			Pose↓			Multi-axis↓		
	Mean		WD_{age}	Mean		WD_{age}	Mean		WD_{age}	Mean		WD_{age}
Flux-Dev	O:90.2	Y:89.4	0.8	O:90.3	Y:88.7	1.6	O:90.3	Y:89.4	0.9	O:90.3	Y:89.1	1.2
Kontext-Pro	Y:26.0	O:23.5	2.4	O:25.6	Y:24.2	1.6	Y:38.3	O:34.9	3.4	Y:32.7	O:32.3	0.8
Nano Banana	Y:34.0	O:32.0	2.0	O:23.3	Y:23.0	0.5	Y:32.5	O:32.1	0.7*	Y:32.0	O:32.0	0.5
Bagel-Edit	Y:27.9	O:25.7	2.3	O:19.3	Y:18.6	1.0	Y:54.1	O:49.6	4.5	Y:34.0	O:32.1	2.0
Qwen-Edit	Y:28.1	O:27.4	0.7	Y:28.0	O:28.0	1.7	Y:48.4	O:45.1	3.4	O:35.5	Y:35.0	1.4

Table 23: Drop_{FSS} % (lower is better) disparity on CelebSET. We reported the mean Drop_{FSS} for Young (Y) and Old (O) groups (left mean value: worst group; right mean value: best group). WD_{age} denotes the Wasserstein distance between the young and old Drop_{FSS} distributions. * marks non-significant values ($p > 0.05$). $WD > 3$ is highlighted.

Model	Accessories↓			Hair↓			Pose↓			Multi-axis↓		
	Mean		WD_{age}	Mean		WD_{age}	Mean		WD_{age}	Mean		WD_{age}
Flux-Dev	O:84.3	Y:83.4	1.5	O:83.8	Y:82.5	2.0	Y:83.2	O:82.8	1.9	O:83.3	Y:82.8	1.2
Kontext-Pro	O:27.9	Y:26.4	2.4	O:21.3	Y:21.3	2.7	Y:31.6	O:29.2	1.7	O:30.7	Y:29.5	3.4
Nano Banana	O:27.1	Y:25.3	3.8	O:17.7	Y:15.6	2.3	O:24.4	Y:21.6	3.2	O:25.1	Y:23.1	3.0
Bagel-Edit	O:28.9	Y:26.8	2.3	O:19.9	Y:17.3	2.8	Y:56.9	O:55.3	2.7	O:34.8	Y:33.8	2.7
Qwen-Edit	Y:48.7	O:44.0	5.1	Y:49.1	O:46.6	3.1	Y:41.8	O:37.7	3.4	Y:46.5	O:43.6	4.2
SeedEdit 3.0	O:27.1	Y:24.6	3.3	O:21.9	Y:17.1	4.9	Y:46.7	O:45.2	2.7	O:33.6	Y:31.2	1.5

Table 24: SC (higher is better) disparity on CelebA. We reported the mean SC for Young (Y) and Old (O) groups (left mean value: worst group; right mean value: best group). WD_{age} denotes the Wasserstein distance between the young and old SC distributions. * marks non-significant values ($p > 0.05$). $WD > 3$ is highlighted. Values scaled by 100.

Model	Accessories↑			Hair↑			Pose↑			Multi-axis↑		
	Mean		WD_{age}	Mean		WD_{age}	Mean		WD_{age}	Mean		WD_{age}
Flux-Dev	O:29.7	Y:35.1	5.4	O:26.9	Y:31.4	4.5	O:31.0	Y:31.5	0.9*	O:27.4	Y:31.2	3.8
Kontext-Pro	Y:73.4	O:73.5	1.0*	Y:75.9	O:76.9	1.1	O:46.6	Y:47.7	1.5*	Y:63.9	O:64.3	0.6*
Nano Banana	O:82.8	Y:85.4	2.6	O:86.9	Y:88.2	1.3	O:62.1	Y:62.8	0.9*	O:73.8	Y:75.5	1.7
Bagel-Edit	O:87.1	Y:88.0	0.9	O:80.2	Y:80.3	0.3*	O:56.8	Y:60.5	3.7	O:68.6	Y:69.0	0.8*
Qwen-Edit	O:79.6	Y:80.0	0.7*	Y:76.5	O:77.0	0.7*	O:65.8	Y:68.1	2.3	O:63.8	Y:65.1	1.2

Table 25: SC (higher is better) disparity on CelebSET. We reported the mean SC for Young (Y) and Old (O) groups (left mean value: worst group; right mean value: best group). WD_{age} denotes the Wasserstein distance between the young and old SC distributions. * marks non-significant values ($p > 0.05$). $WD > 3$ is highlighted. Values scaled by 100.

Model	Accessories↑			Hair↑			Pose↑			Multi-axis↑		
	Mean		WD_{age}	Mean		WD_{age}	Mean		WD_{age}	Mean		WD_{age}
Flux-Dev	O:26.4	Y:30.3	3.9	O:27.3	Y:28.8	1.6*	O:28.5	Y:29.0	1.4*	O:26.4	Y:27.4	1.3*
Kontext-Pro	Y:72.5	O:73.5	1.6*	O:74.5	Y:75.0	0.9*	O:35.4	Y:40.5	5.1	O:59.9	Y:61.2	1.5*
Nano Banana	Y:82.1	O:84.7	2.6	O:84.6	Y:86.8	2.2	O:58.7	Y:59.5	1.1*	O:72.1	Y:73.4	1.5*
Bagel-Edit	Y:86.2	O:88.5	2.2	Y:77.4	O:78.3	1.1*	Y:59.9	O:61.1	1.5*	O:67.5	Y:68.4	1.3*
Qwen-Edit	O:81.9	Y:82.1	0.8*	Y:77.2	O:78.7	1.5*	O:60.8	Y:66.1	5.4	O:65.0	Y:66.4	1.7*
SeedEdit 3.0	Y:86.0	O:87.1	1.1*	O:80.9	Y:82.2	1.5*	O:63.7	Y:64.2	2.9*	Y:69.1	O:69.5	1.3*

Table 26: OER (lower is better) disparity on CelebA. We reported the mean OER for the Young (Y) and Old (O) groups (left mean value: worst group; right mean value: best group). WD_{age} denotes the Wasserstein distance between the young and old OER distributions. * marks non-significant values ($p > 0.05$). $WD > 3$ is highlighted. Values scaled by 100.

Model	Accessories↓			Hair↓			Pose↓			Multi-axis↓		
	Mean		WD_{age}	Mean		WD_{age}	Mean		WD_{age}	Mean		WD_{age}
Flux-Dev	O:78.6	Y:65.0	13.6	O:78.8	Y:62.3	16.5	O:59.6	Y:43.9	15.7	O:79.0	Y:60.8	18.3
Kontext-Pro	O:14.6	Y:10.3	4.3	O:24.8	Y:17.8	7.0	O:18.9	Y:15.7	3.4	O:28.0	Y:21.1	6.9
Nano Banana	O:14.1	Y:9.2	4.9	O:20.0	Y:13.4	6.7	O:13.5	Y:9.1	4.5	O:23.6	Y:15.6	8.0
Bagel-Edit	O:11.1	Y:7.6	3.7	O:24.4	Y:16.5	7.9	O:16.3	Y:13.0	3.6	O:25.9	Y:19.4	6.5
Qwen-Edit	O:16.6	Y:12.7	3.9	O:30.0	Y:23.2	6.9	O:20.4	Y:19.7	1.0*	O:30.9	Y:26.1	4.8

Table 27: OER (lower is better) disparity on CelebSET. We reported the mean OER for Young (age: 20 – 30) and Old (age: 50 – 60) groups (left mean value: worst group; right mean value: best group). WD_{age} denotes the Wasserstein distance between the young and old OER distributions. * marks non-significant values ($p > 0.05$). $WD > 3$ is highlighted. Values scaled by 100.

Model	Accessories↓			Hair↓			Pose↓			Multi-axis↓		
	Mean		WD_{age}	Mean		WD_{age}	Mean		WD_{age}	Mean		WD_{age}
Flux-Dev	O:81.3	Y:58.6	22.9	O:80.6	Y:58.5	22.1	O:62.6	Y:37.3	25.3	O:80.1	Y:56.5	23.8
Kontext-Pro	O:14.8	Y:13.0	3.1	O:26.3	Y:20.9	5.7	O:14.7	Y:13.7	2.4*	O:29.9	Y:22.8	7.3
Nano Banana	Y:21.6	O:15.4	7.4	O:21.8	Y:12.5	9.2	O:9.2	Y:6.9	2.5*	O:29.5	Y:24.1	10.1
Bagel-Edit	Y:12.7	O:11.8	5.1	O:26.2	Y:25.9	7.7	O:13.1	Y:12.5	2.3*	O:34.5	Y:20.1	14.1
Qwen-Edit	O:25.0	Y:21.9	3.1*	O:36.0	Y:29.9	6.4	O:23.4	Y:13.5	9.9	O:39.3	Y:29.6	9.8
SeedEdit 3.0	O:10.9	Y:7.5	3.4	O:25.7	Y:15.8	9.9	O:15.4	Y:14.3	2.9*	O:25.3	Y:26.9	5.7

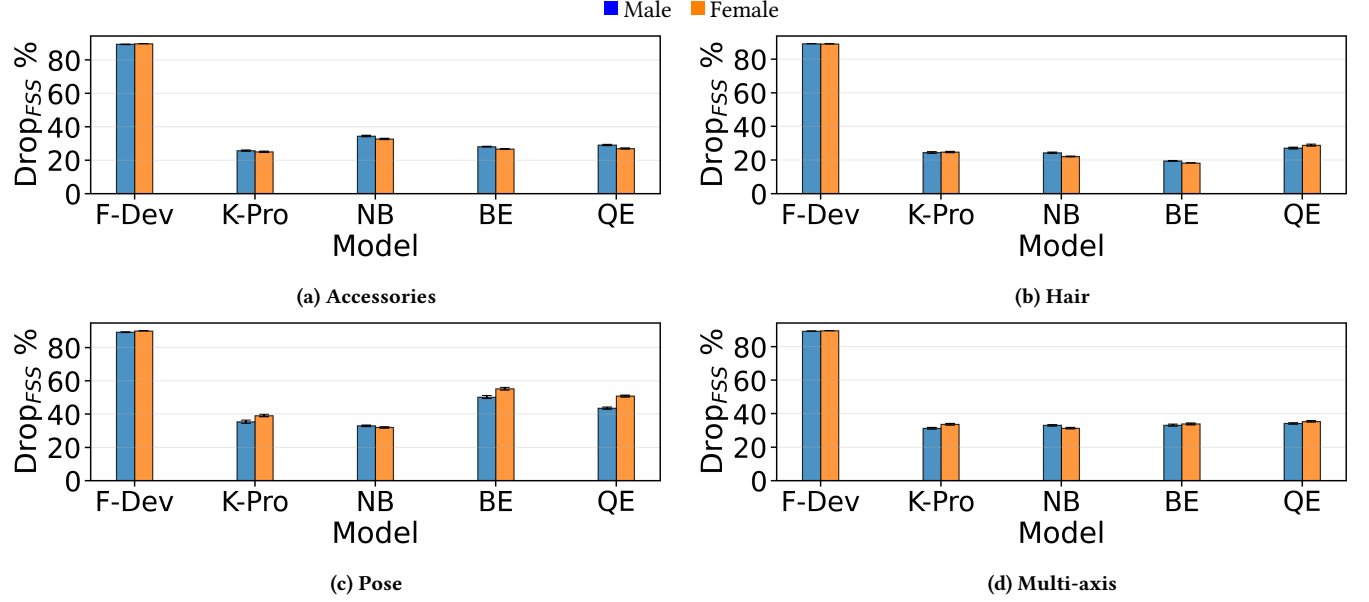


Figure 6: Drop_{FSS} % (lower is better) across gender subgroups in CelebA. Flux-Dev (F-Dev), Kontext-Pro (K-Pro), Nano Banana (NB), Bagel-Edit (BE), and Qwen-Edit (QE).

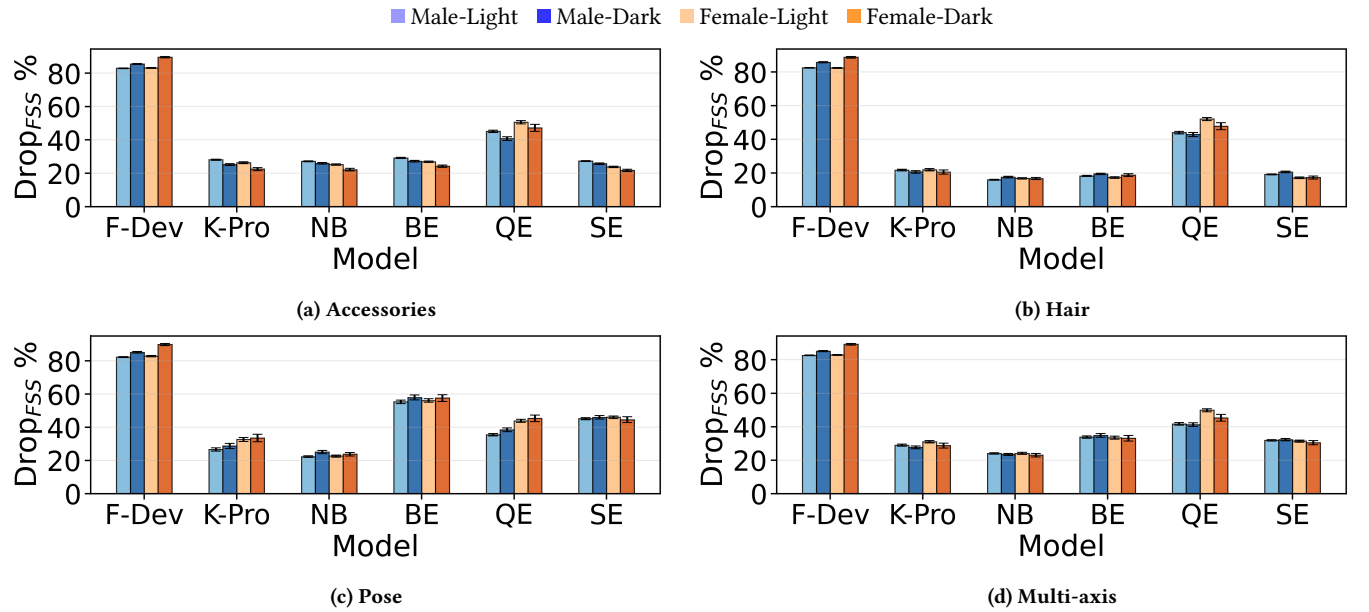
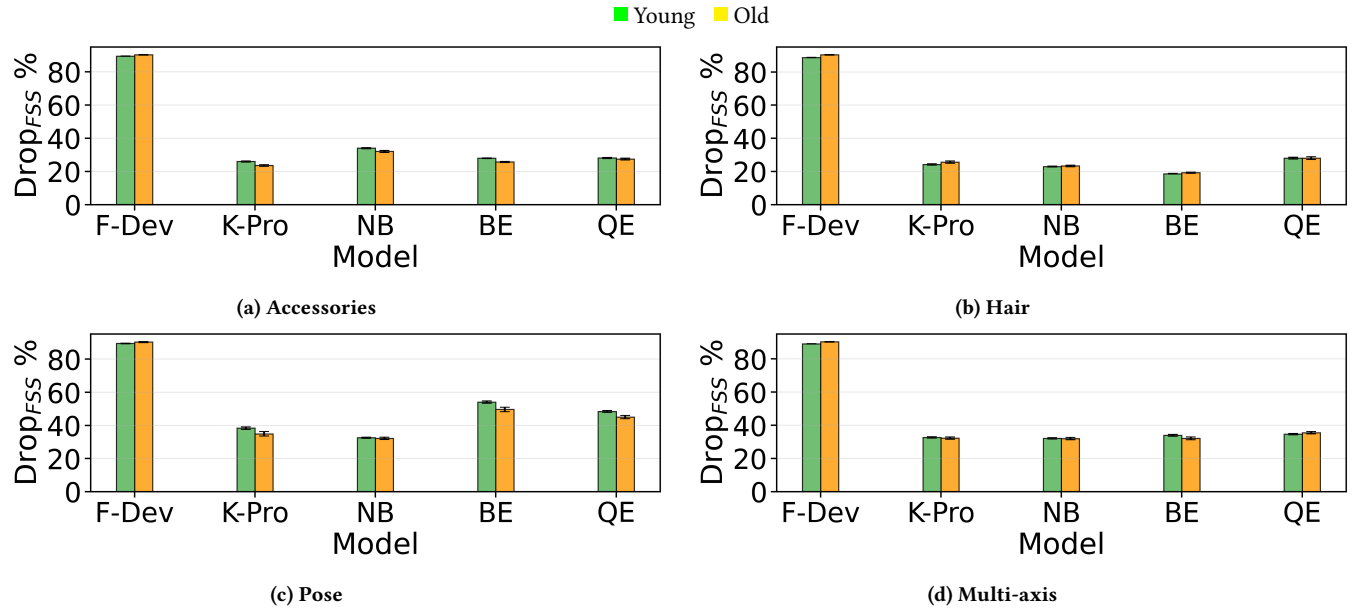
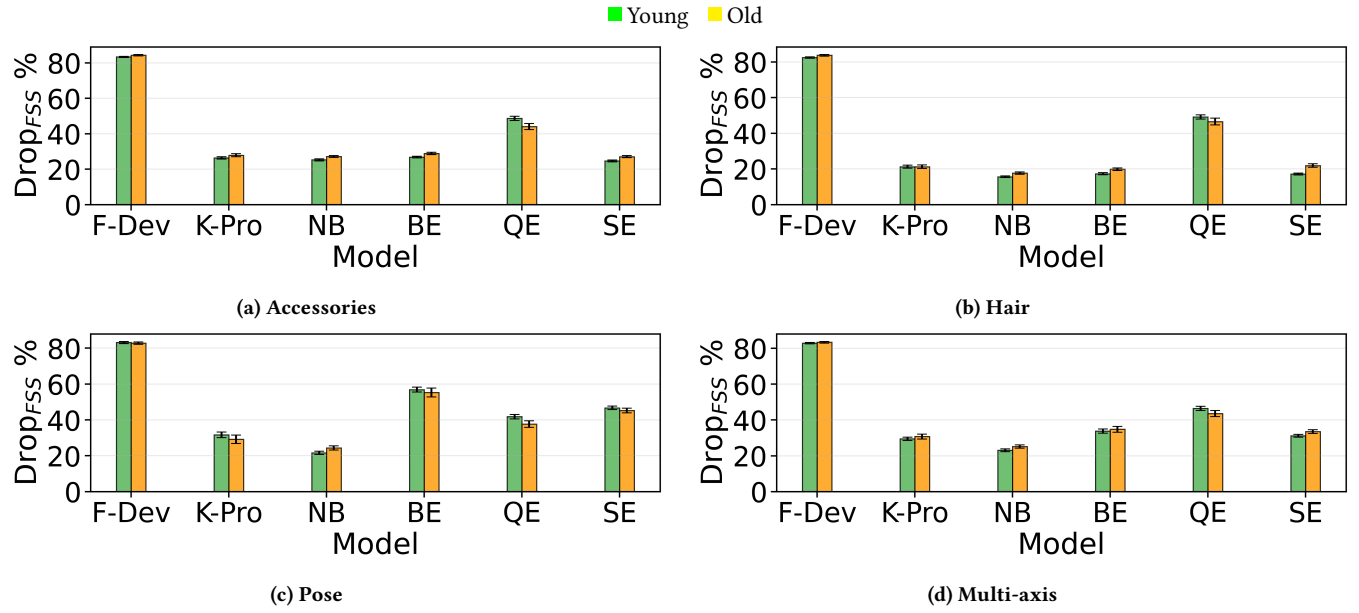


Figure 7: Drop_{FSS} % (lower is better) across intersectional demographic subgroups in CelebSET.

Figure 8: Drop_{FSS} % (lower is better) across age subgroups in CelebA.Figure 9: Drop_{FSS} % (lower is better) across age subgroups in CelebSET.

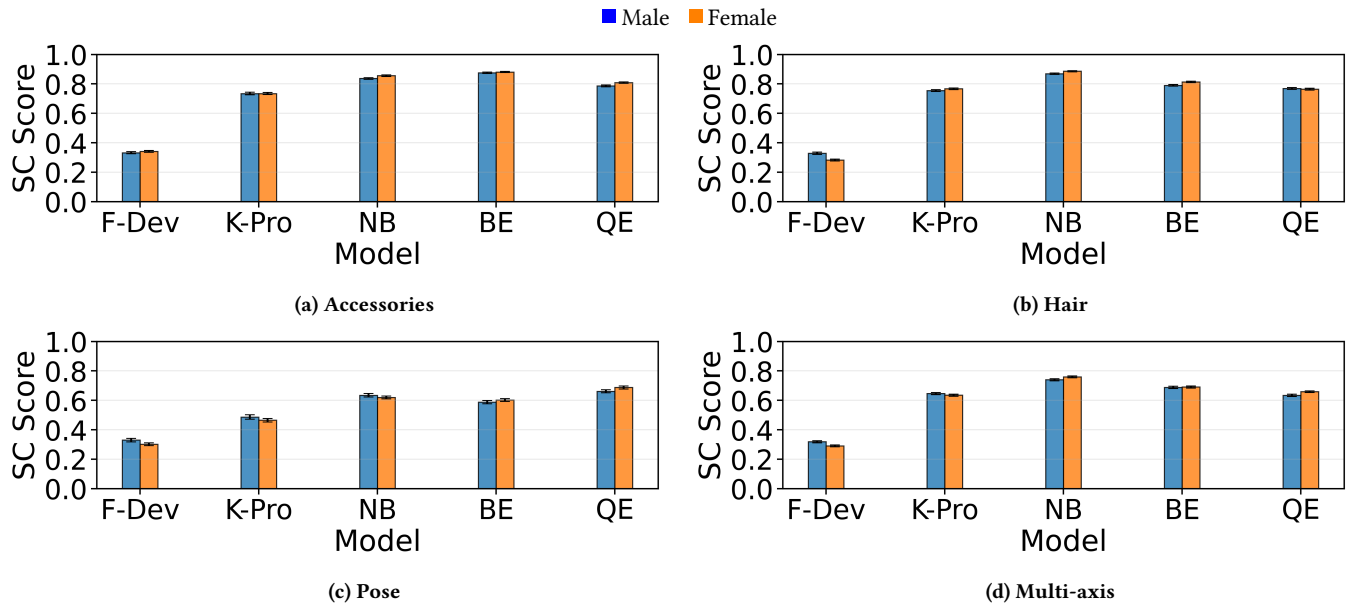


Figure 10: SC (higher is better) across gender subgroups in CelebA.

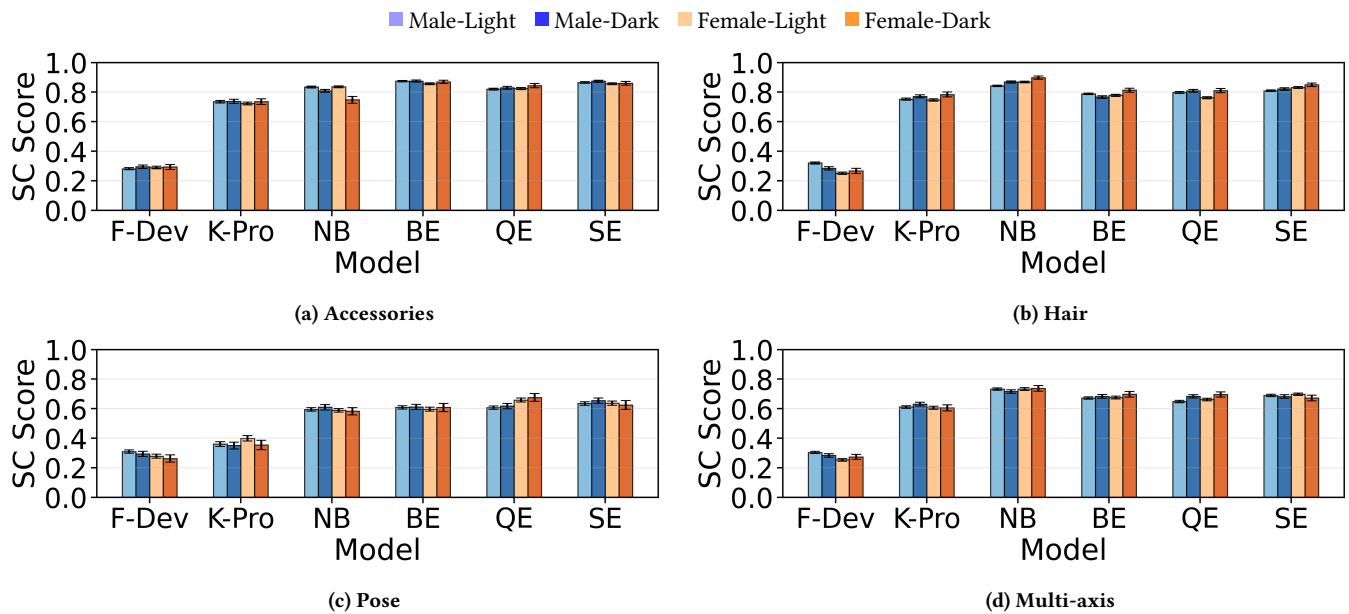


Figure 11: SC (higher is better) across intersectional demographic subgroups in CelebSET.

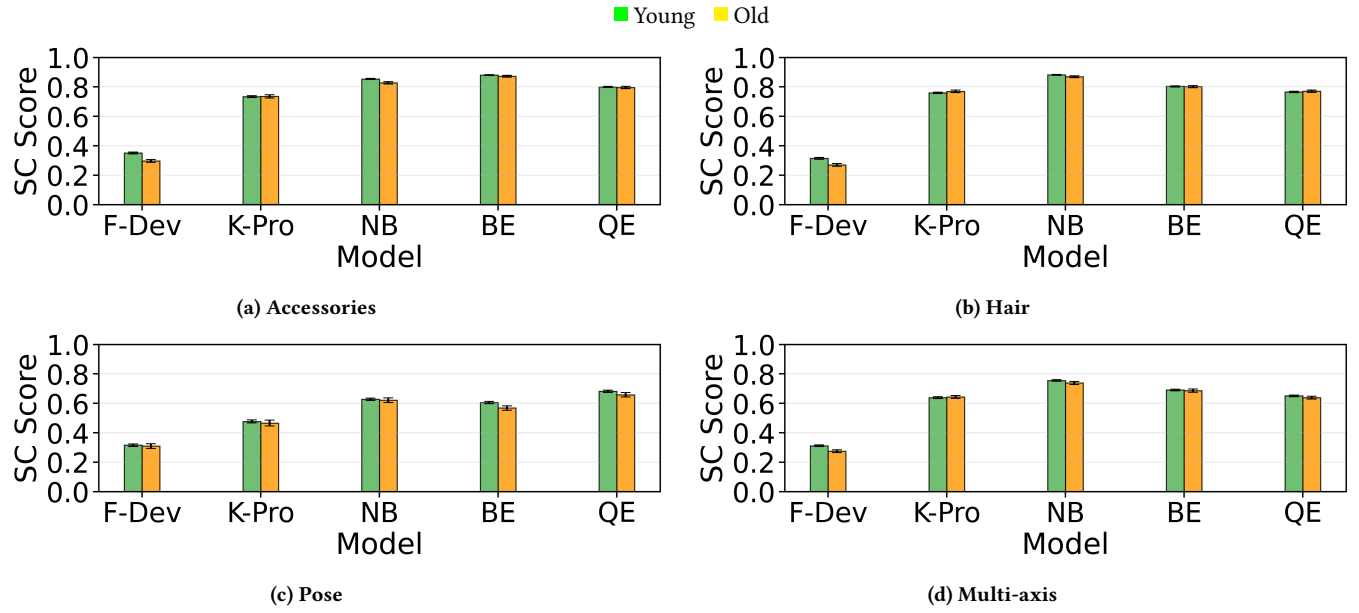


Figure 12: SC (higher is better) across age subgroups in CelebA.

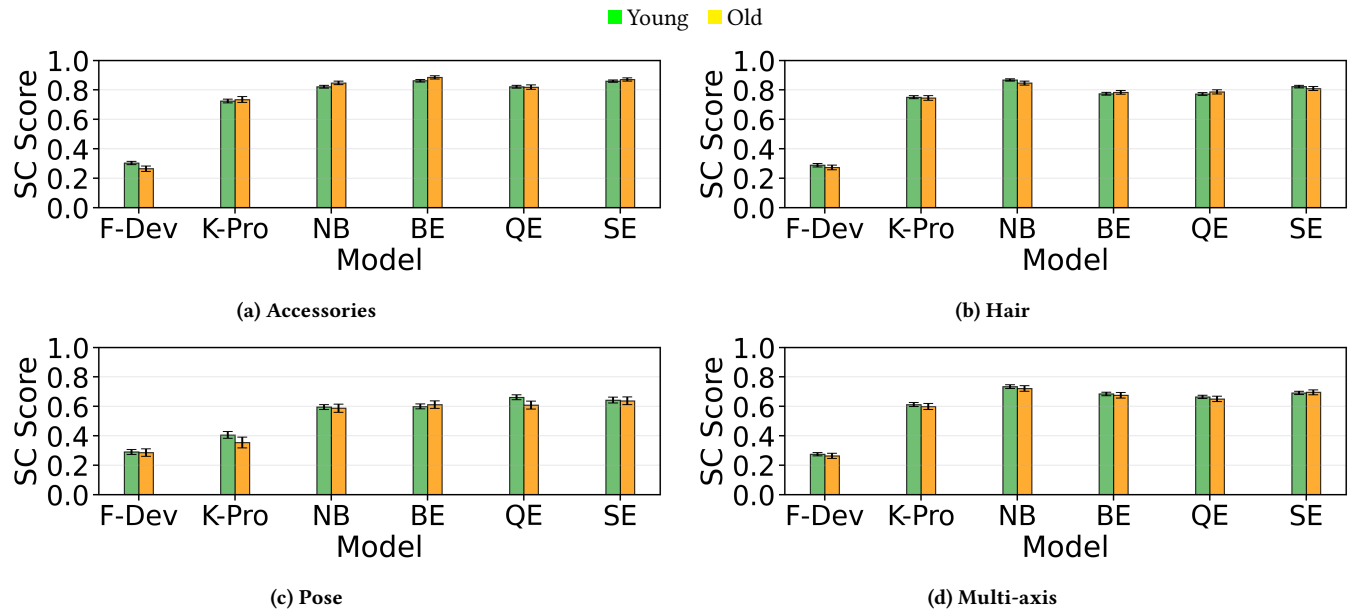


Figure 13: SC (higher is better) across age subgroups in CelebSET.

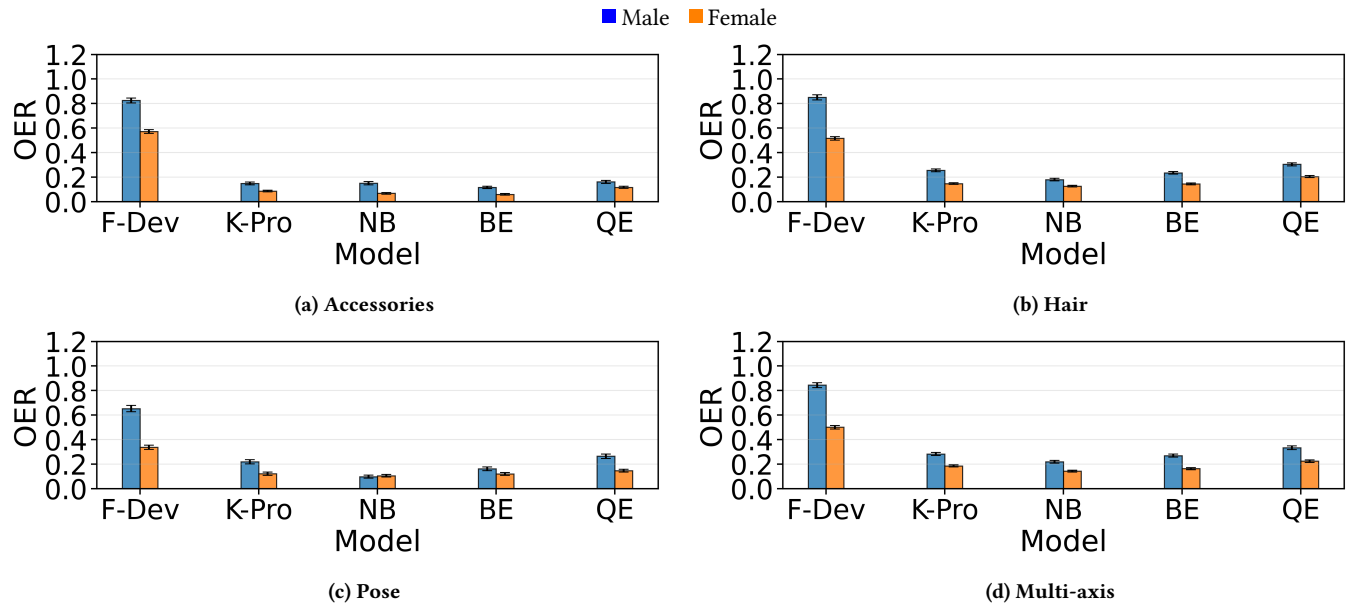


Figure 14: OER (lower is better) across gender subgroups in CelebA.

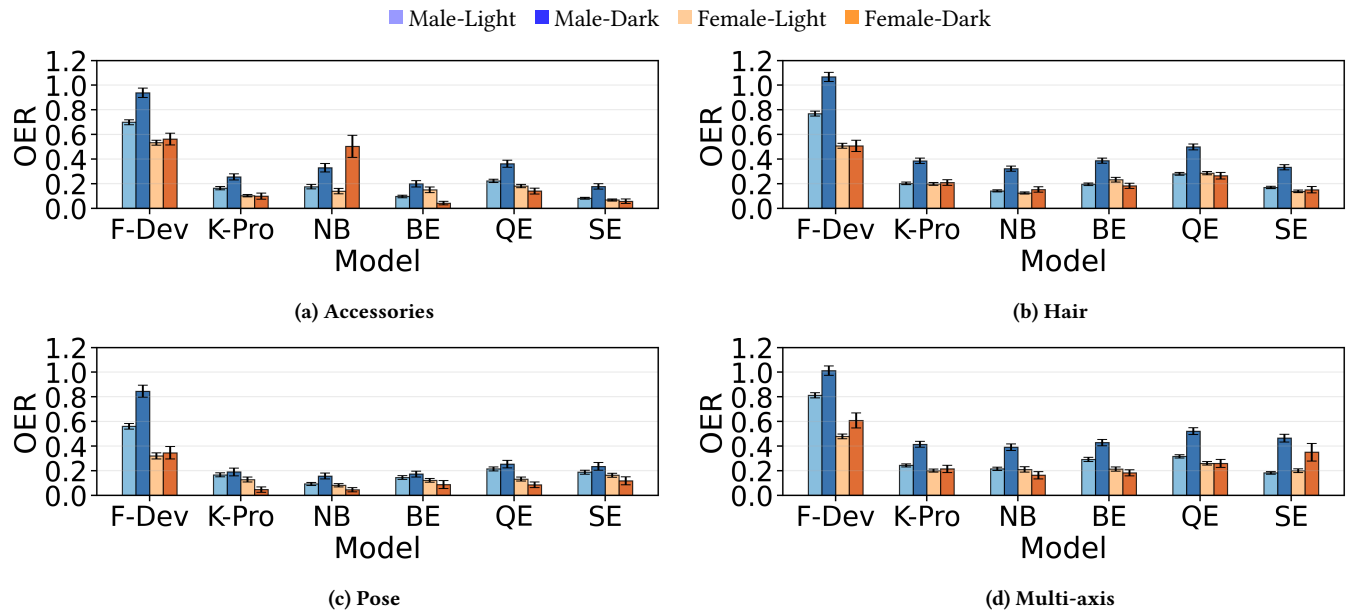


Figure 15: OER (lower is better) across intersectional demographic subgroups in CelebSET.

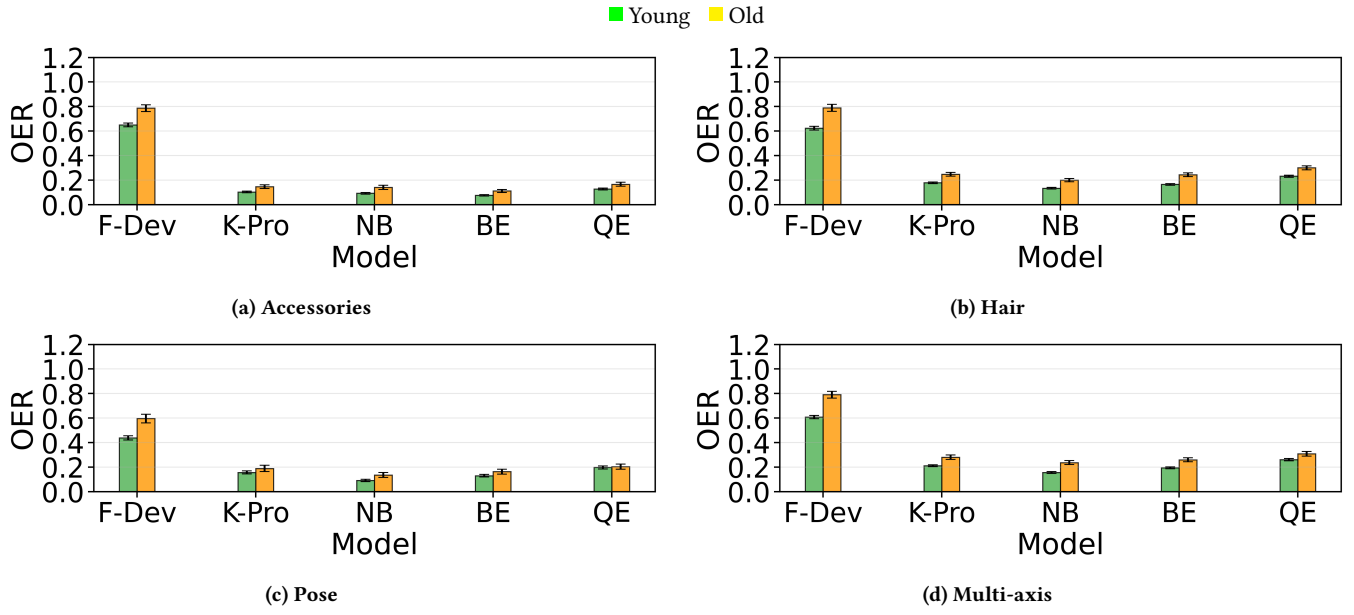


Figure 16: OER (lower is better) across age subgroups in CelebA.

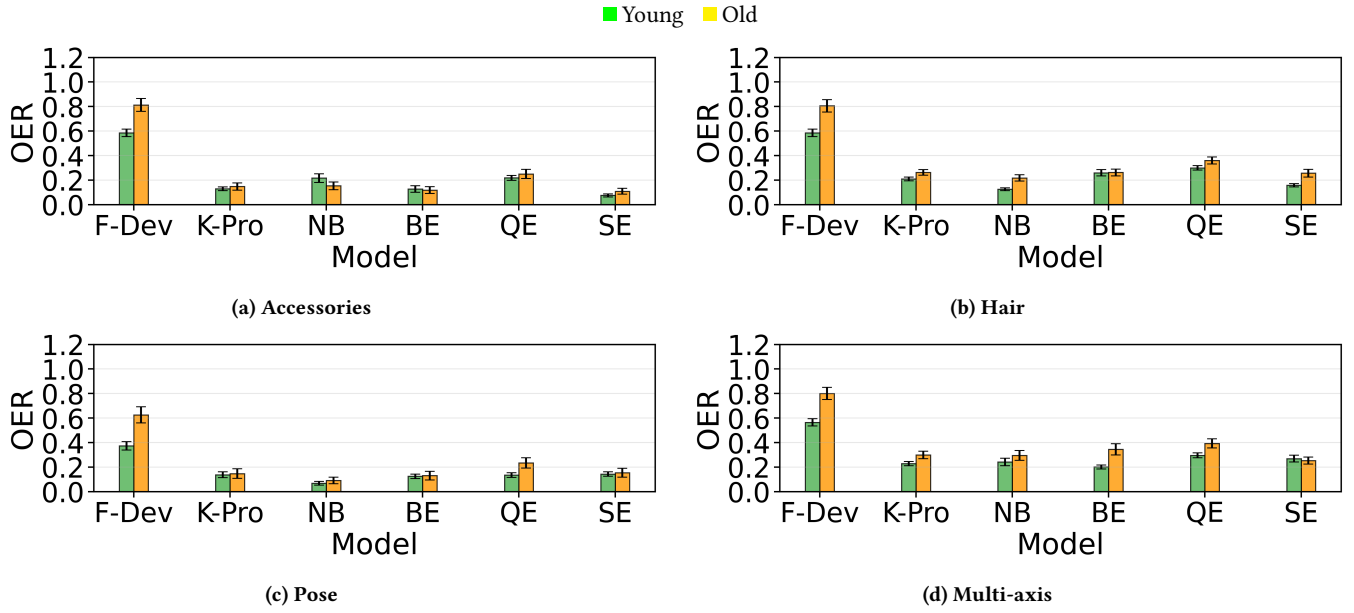


Figure 17: OER (lower is better) across age subgroups in CelebSET.

Table 28: Attributes that frequently caused demographic biases in overediting. Model abbreviations: Nano-Banana (NB), SeedEdit (SE), Bagel-Edit (BE), Qwen-Edit (QE), Flux-Dev (F-Dev), and Kontext-Pro (K-Pro). Demographic abbreviations: Male (M), Female (F), Old (O), Young (Y), Light skin tone (L), and Dark skin tone (D).

Gender Bias — CelebA			Gender Bias — CelebSET		
Attribute	# Models	Model Names	Attribute	# Models	Model Names
spiked cut	5 (4M, 1F)	F-Dev, K-Pro, NB, BE, QE	curly hair	6 (6M)	F-Dev, K-Pro, NB, BE, QE, SE
straight hair	5 (5M)	F-Dev, K-Pro, NB, BE, QE	fumanchu mustache	6 (6M)	F-Dev, K-Pro, NB, BE, QE, SE
curly hair	4 (4M)	F-Dev, NB, BE, QE	brown hair	6 (6M)	F-Dev, K-Pro, NB, BE, QE, SE
fumanchu mustache	3 (3M)	F-Dev, K-Pro, NB	straight hair	6 (6M)	F-Dev, K-Pro, NB, BE, QE, SE
blue hair	3 (3M)	F-Dev, K-Pro, QE	beanie hat	5 (5M)	K-Pro, NB, BE, QE, SE
mutton chops	3 (1M, 2F)	K-Pro, NB, BE	cowboy hat	5 (5M)	K-Pro, NB, BE, QE, SE
bob cut	3 (3M)	K-Pro, NB, QE	toothbrush mustache	4 (4M)	F-Dev, K-Pro, BE, SE
green hair	3 (3M)	K-Pro, BE, QE	spiked cut	4 (4M)	K-Pro, BE, QE, SE
black hair	3 (3M)	K-Pro, NB, BE	bob cut	3 (3M)	F-Dev, K-Pro, QE
brown hair	3 (3M)	K-Pro, BE, QE	mutton chops	3 (1M, 2F)	K-Pro, NB, SE

Age Bias — CelebA			Age Bias — CelebSET		
Attribute	# Models	Model Names	Attribute	# Models	Model Names
curly hair	5 (5O)	F-Dev, K-Pro, NB, BE, QE	curly hair	6 (6O)	F-Dev, K-Pro, NB, BE, QE, SE
bob cut	5 (5O)	F-Dev, K-Pro, NB, BE, QE	spiked cut	6 (6O)	F-Dev, K-Pro, NB, BE, QE, SE
spiked cut	4 (4O)	F-Dev, K-Pro, BE, QE	bob cut	5 (5O)	F-Dev, K-Pro, BE, QE, SE
green hair	4 (4O)	F-Dev, K-Pro, BE, QE	straight hair	5 (5O)	K-Pro, NB, BE, QE, SE
straight hair	4 (4O)	K-Pro, NB, BE, QE	brown hair	4 (4O)	F-Dev, K-Pro, QE, SE
brown hair	4 (4O)	K-Pro, NB, BE, QE	beret hat	3 (1O, 2Y)	F-Dev, NB, QE
blue hair	3 (3O)	F-Dev, K-Pro, BE	beanie hat	3 (3O)	F-Dev, K-Pro, NB
full beard	3 (3O)	K-Pro, NB, QE	blue hair	3 (3O)	K-Pro, NB, SE
cowboy hat	3 (3O)	K-Pro, NB, BE	pencil mustache	3 (2O, 1Y)	NB, BE, SE
black hair	3 (3O)	K-Pro, NB, BE	fumanchu mustache	2 (2O)	F-Dev, QE

Skin Tone Bias — CelebSET		
Attribute	# Models	Model Names
fumanchu mustache	6 (6D)	F-Dev, K-Pro, NB, BE, QE, SE
curly hair	6 (6D)	F-Dev, K-Pro, NB, BE, QE, SE
brown hair	6 (6D)	F-Dev, K-Pro, NB, BE, QE, SE
straight hair	5 (5D)	K-Pro, NB, BE, QE, SE
beanie hat	5 (5D)	K-Pro, NB, BE, QE, SE
spiked cut	4 (4D)	F-Dev, K-Pro, NB, SE
green hair	4 (4D)	F-Dev, K-Pro, BE, QE
beret hat	3 (3D)	F-Dev, NB, BE
cowboy hat	3 (3D)	K-Pro, NB, BE
black hair	3 (3D)	BE, QE, SE

J Which editing attributes commonly led to demographic biases in overediting?

Across our four evaluation sets (hair, accessories, pose, and multi-axis), we used 40 unique editing attributes in this study (e.g., curly hair, handlebar mustache). We identified which of these 40 attributes commonly contributed to demographic bias in overediting. For each {dataset, bias axis} combination (e.g., CelebA, gender), we computed WD using the OER metric (WD_g for this example) for all 40 attributes, across all editing models.

For each {dataset, bias axis} combination, we identified attributes that frequently occurred in the top-10 WD scores (highest overediting bias) across all models. We reported our results in Table 28. In each table, we showed: (1) editing attribute (e.g., curly hair), (2) the number of models in which that attribute came in the top-10 WD scores. We also showed the bias direction in parentheses. For instance, (4M, 1F) means that in four models, males showed more overedits and in one model, females showed more overedits, (3) the names of all those models. We reported 2 tables for CelebA (gender, age), and 3 tables for CelebSET (gender, skin-tone, and age).

Across all five tables, we see a common trend: in almost all models, the highest demographic biases in overediting are caused by hair-based editing operations. This includes almost all types of hair edit operations like hair style (e.g., spiked cut), facial hair (e.g., fumanchu mustache), and hair color (e.g., green hair). Across the five tables, results are consistent with Section 3.0.2: models showed more overedits on men, dark-skinned faces, and old faces.

K FSS: Demographic Bias Analysis on Base Images

To check if FSS showed demographic bias in the base images, we computed the average FSS using the base images for each demographic group. To measure disparity (or bias) between groups, we used the WD metric. For each group, we also reported mean values (with 95% confidence intervals). For CelebA, we showed gender comparisons in Table 29a and age comparisons in Table 29b. For CelebSET, we showed gender-skin-tone comparisons in Table 29c and age comparisons in Table 29d. Across all tables, we observed no significant demographic bias (WD always < 3).

L FSS: Occlusion Sensitivity Analysis on Base Images

We used CelebA base images because only CelebA has annotations for facial accessories such as glasses and hats. We created two groups: “with accessory” (faces with hats, glasses, or both: 34 images), and “without accessory” (515 images). Mean FSS for “with accessory”: 96.77 ± 0.37 , “without accessory”: 96.82 ± 0.10 .

WD between these groups: 0.14 (WD much less than 3), with a p -val = 0.9 (WD is not statistically significant). FSS is not sensitive to occlusions: scores are similar even when base images include accessories such as glasses or hats.

M Stereotype Analysis

We conducted a controlled experiment to check if models performed edits better on male faces, if editing instructions are male-centric (e.g., add a handlebar mustache). From the 140 instructions that we used for evaluation, we selected 17 instructions that generally apply to males (e.g., add a pencil mustache, add a bushy beard). All of these instructions focused on editing facial hair.

For this subset of instructions (called the male-centric instruction subset), we computed the average SC across all models, for male and female groups in CelebA, and for four intersectional subgroups (male-light, female-light, male-dark, female-dark) in CelebSET. CelebA results are shown in Table 30a, and CelebSET results are shown in Table 30b. In the “Mean” column for CelebA (Table 30a), the left value is the worst group, and the right value is the best group. For CelebSET (Table 30b), we reported two worst groups (out of four intersectional groups): the left value is the worst, and the right is the second-worst performing group. For CelebA, we reported WD_g and for CelebSET, we reported both WD_g and WD_s .

In CelebA (Table 30a), we saw gender bias in three models. Out of these, only two models showed worse SC on females (Kontext-Pro and Bagel-Edit). In CelebSET (Table 30b), only for Flux-Dev, the worst and the second-worst groups were both females (FL and FD). For all other models, at least one male intersectional group was present in the two worst groups.

For male-centric editing instructions, we did not find clear trends indicating that models perform edits better for males.

N Face ID Progression Analysis

We created line plots (in Figure 18) to assess how facial similarity scores changed in models after each edit. In the x-axis, we marked the editing steps, starting from the base image (edit 0) to the final edited image (edit 4). At each edit step, the y-axis is the average facial similarity score (FSS) of a model (with 95% confidence intervals) for all images in that edit step. The left plot in Figure 18 shows FSS when base images are from CelebA, while the right plot shows results on CelebSET. In both these plots, we showed the aggregated FSS across all four test sets (hair, accessories, pose, and multi-axis).

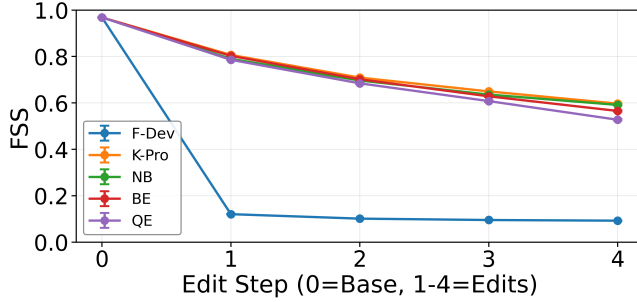
For all models, the facial similarity score (FSS) consistently dropped with every edit. In Figure 18, for both CelebA and CelebSET, we see the biggest drop in FSS from edit 0 to 1. This is expected, as in edit 1, we transition from a real image to a model-edited image. In subsequent edits, we see smaller (but consistent) drops.

Table 29: Baseline FSS (%) across demographic groups on CelebA and CelebSET. Subscripts denote 95% CI half-widths. * superscript indicates a non-significant p-value ($p > 0.05$). ML: Male-Light, FL: Female-Light, MD: Male-Dark, FD: Female-Dark.

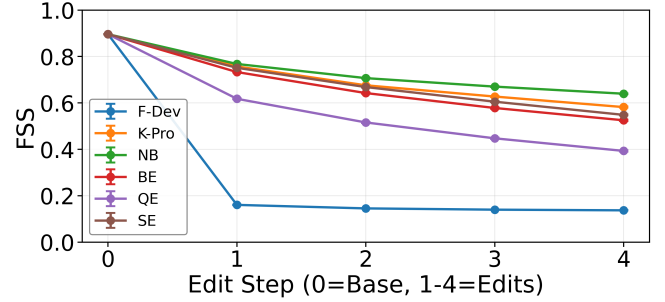
(a) CelebA - Gender			(b) CelebA - Age			(c) CelebSet - Gender & Skin			(d) CelebSet - Age		
Group	Mean _{CI}	Count	Group	Mean _{CI}	Count	Group	Mean _{CI}	Count	Group	Mean _{CI}	Count
Male	96.79 _{0.15}	244	Young	96.79 _{0.11}	413	ML	89.26 _{0.51}	222	Young	90.14 _{0.91}	98
Female	96.85 _{0.14}	305	Old	96.91 _{0.18}	136	MD	89.59 _{0.85}	95	Old	88.47 _{0.89}	41
WD _g ↓	0.13*		WD _{age} ↓	0.28		FL	90.06 _{0.69}	161	WD _{age} ↓	1.74	
						FD	89.67 _{1.22}	39			
						WD _g ↓	0.63*				
						WD _s ↓	0.30*				

Table 30: Stereotype analysis using SC. For the male-centric instruction subset, we reported the mean SC for demographic groups. In the Mean column, for CelebA: left is the worst group; right value is the best group. For CelebSET: left: worst group; right: second-worst group. * marks non-significant values ($p > 0.05$). WD > 3 is highlighted. Values scaled by 100.

(a) CelebA					(b) CelebSET				
Model	Male-Centric Subset [†]				Model	Male-Centric Subset [†]			
	Mean		WD _g			Mean	WD _g	WD _s	
Flux-Dev	F:25.4	M:31.2	5.8		Flux-Dev	FL:24.8	FD:25.5	9.4	0.9*
Kontext-Pro	F:54.1	F:55.9	1.8		Kontext-Pro	FL:54.0	ML:55.9	2.1	2.1
Nano Banana	M:70.7	F:74.2	4.4		Nano Banana	MD:66.1	ML:72.0	5.3	4.8
Bagel-Edit	F:58.2	M:61.4	3.2		Bagel-Edit	FL:56.8	MD:61.0	3.4	1.8*
Qwen-Edit	F:39.3	M:39.8	1.1*		Qwen-Edit	FL:43.1	ML:44.5	3.6	5.8
					SeedEdit 3.0	ML:43.4	FL:44.5	1.9	2.3



(a) CelebA



(b) CelebSET

Figure 18: Facial similarity scores after each edit. The x-axis indicates the editing step, from 0 (base image) to 4 (final edited image). At each edit step, the Y-axis is the average facial similarity of a model for all images in that edit step. Note that 95% confidence intervals are added, but are too narrow to be visible.