Mitigating the Impact of Attribute Editing on Face Recognition

Sudipta Banerjee, Sai Pranaswi Mullangi, Shruti Wagle, Chinmay Hegde, Nasir Memon
New York University
{sb9084, sm11006, sgw6735, chinmay.h, memon}@nyu.edu

Abstract

Through a large-scale study over diverse face images, we show that facial attribute editing using modern generative AI models can severely degrade automated face recognition systems. This degradation persists even with identity-preserving generative models. To mitigate this issue, we propose two novel techniques for local and global attribute editing. We empirically ablate twenty-six facial semantic, demographic and expression-based attributes that have been edited using state-of-the-art generative models, and evaluate them using ArcFace and AdaFace matchers on CelebA, CelebAMaskHQ and LFW datasets. Finally, we use LLaVA, an emerging visual question-answering framework for attribute prediction to validate our editing techniques. Our methods outperform the current state-of-the-art at facial editing (BLIP, InstantID) while improving identity retention by a significant extent.

1. Introduction

Motivation: Facial attribute editing refers to the process of digitally altering semantic cues to change the visual appearance in a face image. Examples may include changing hair color, adding accessories such as eyeglasses, and altering facial expressions; all these go considerably beyond digital re-touching [3, 5] via smartphone applications such as YouCam Makeup and Facetune Editor. Generative AI-based tools for attribute modification are readily available; many are based on conditional generative models (examples include Cafe-GAN, STGAN, GAN-Edit, Mask-aware editing, Age-GAN, and Talk-to-Edit) [21, 25, 26, 36, 41, 18].

Although attribute editing models are typically guided by aesthetic or creative pursuits, their impact on face recognition (FR) is relatively under-explored [1]. Automated face matchers demonstrate bias towards semantic attributes and accessories as shown in [38]. Additionally, if digital manipulations are induced, that may adversely impact face recognition. Can generative models be therefore used to deliberately perturb the *biometric* quality of an image? This question is paramount from the security perspective if digital attribute editing can potentially be used as a passive software-based presentation attack with the intent of evasion (conceal one's true identity) [19, 29]. The recent, massive proliferation of open-source generative diffusion models capable of face synthesis and editing adds urgency to this question [9].

Objective: Our goal will be to investigate the impact of a diverse set of semantic, demographic and expression-based attribute editing operations using generative models on automated FR systems. Most generative models rely on MSE, FID or SNR [7] to preserve perceptual quality and not biometric utility. Existing FR systems are equipped to handle low resolution, pose and demographic variations [12, 20] but may falter in presence of digitally induced perturbations. So, we propose techniques that can mitigate the impact of digital attribute editing on FR.

Our Approach: Facial attribute editing is a complex task that can require either minor adaptations locally (e.g., changing hair color) or major alterations globally (e.g., changing age). In both cases, attribute editing should not compromise biometric fidelity. Therefore, we formulate the problem of attribute editing at two levels: (1) *local* attribute editing, that changes fine facial details. To achieve this, we use segmentation masks to perform the desired edit and depth maps to preserve the extraneous details excluded from the mask. We only need a single facial image to perform local editing. (2) *global* attribute editing, that changes the coarse facial details. To achieve this, we use a regularization method supervised by identity-preserving loss and utilizes few exemplars of the desired attribute (30 images) and the target individual (10 images). We explore a suite of generative models operating in text-to-image (txt2img) and image-to-image (img2img) modes to conduct our investigation. Table 1 outlines our different methods. See Fig. 1 for the overview of the proposed framework.
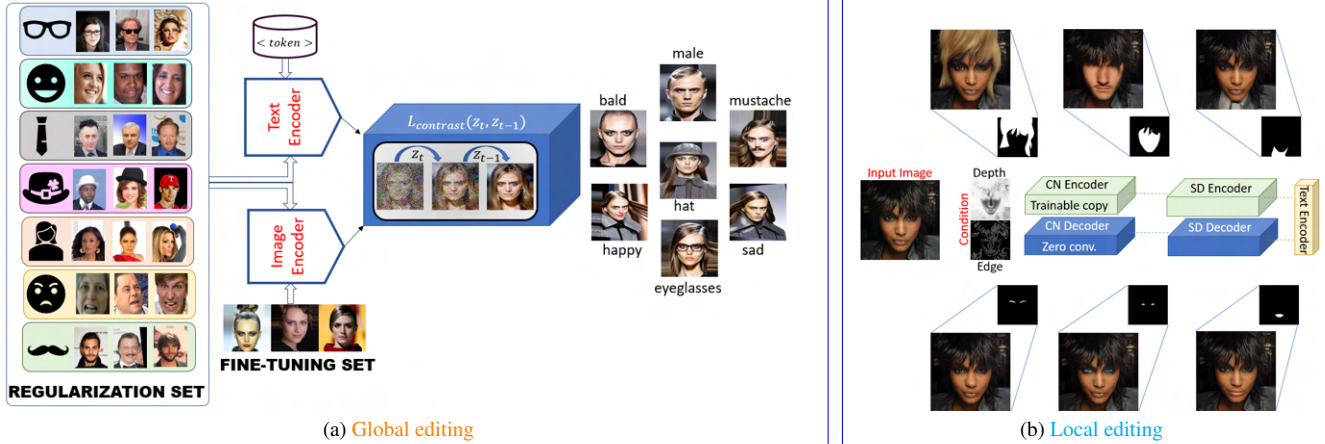
(a) Global editing   (b) Local editing

Figure 1. **(a)** The *global* editing framework uses a pre-trained latent diffusion model (DreamBooth) and fine-tunes it with few samples of a target individual and a regularization set containing image-caption pairs belonging to different attributes. The model uses contrastive learning to generate attribute-edited images while maintaining biometric fidelity. This method operates in *txt2img* mode. **(b)** The *local editing* framework uses a pre-trained latent diffusion model (SD) along with Control Net (CN) for inpainting guided via segmentation masks. CN uses conditional input such as depth map or edge map for detail preservation and mask for attribute editing. Note white pixels in the mask indicate the local regions to be edited while preserving fidelity with the reference input image. This method operates in both *txt2img* and *img2img* modes. Both methods use a text prompt to edit the input with desired attribute.

Contributions: (i) We conduct a comprehensive analysis of the effect of facial attribute editing via diffusion models on FR. We study 3 attribute classes: (a) facial semantic attributes (hair style, hair color, facial hair, structure and accessories), (b) demographic attributes (age and sex), and (c) facial expressions (happy, fear, neutral, surprise, disgust, sad and anger), resulting in a total of 26 attributes; additionally, we apply a reconstruction operation, called, 'No attrib' that outputs the original image without attribute editing.
(ii) We propose mitigation techniques that integrate regularization-based techniques supervised by loss functions (contrastive, cosine embedding and smooth-L_1) for global attribute editing, and regularization-free methods guided by conditional inputs (depth map, edge map and semantic segmentation mask) for local attribute editing.
(iii) We use an open-ended VQA model based on LLaVA to automatically assess the correctness of the desired attribute editing operation. This is a significant deviation from prior work that typically rely on user studies on a limited scale. Automated binary classifiers can be used alternatively but is not scalable for unseen or multi-attribute editing.

2. Related Work

Digital retouching [3] and facial attribute editing using generative adversarial networks (GANs) [14, 25, 21, 26, 41, 5, 37, 36, 22, 18, 6] are highly common. Attribute editing is typically used for creative pursuits and the edited images are evaluated in terms of perceptual quality metrics such as FID and IS. In [1], the authors showed that GAN-based facial attribute editing can significantly degrade automated face recognition performance although the images may be visually perceptible. Image personalization has significantly evolved after the introduction of diffusion models that require a simple text prompt and some images of the target individual (in some cases, even a single image will suffice). The highly non-linear operations performed by the diffusion models may degrade the biometric utility of the generated images, specifically, after semantic editing. Bias in face images generated by diffusion models focus mostly on demographic attributes (age, sex and ethnicity) [2, 30]. Some work propose prompt engineering for de-biasing [8]. ChatFace [42] performs text-driven facial image editing in the semantic latent space of diffusion model This system leverages LLMs to perform complex multi-attribute manipulations through dialogue, similar to Talk-to-Edit [18] which uses interactive facial editing via GANs.

In this work, we use state-of-the-art baselines involving subject driven semantic editing: DreamBooth with prior preservation (DB-base) [33], InstantID [39] and BLIP-Diffusion [23]. InstantID is a tuning-free method to achieve ID-preserving generation with only single image that focuses on the detected face and performs novel view synthesis, stylized generation, identity interpolation, etc. BLIPDiffusion introduces a multi-modal encoder which is pre-trained to provide subject representation, called prompted context generation that renders the learned subject in new contexts. BLIPDiffusion requires a reference image for transferring the style to the input. We are unable to compare with ChatFace [42] as the codes are not

Table 1. Taxonomy of methods explored in this work. The top four methods achieve global editing; the bottom three achieve local editing. The methods shaded in blue perform the best in terms of biometric preservation and successful attribute editing.

Method	Conversion Type	Model Details	#Input Images	Conditional Input
DB-base: DreamBooth (with prior loss)	*txt2img*	SD V1.5 SD V2.0	10	Reg Set
DB-prop: DreamBooth (with prior + contrastive loss)	*txt2img*	SD V1.5	10	Reg Set
TI: Textual Inversion (with MSE loss)	*txt2img*	SD V1.5	10	No. of TI vectors
TI$_{cs}$: Textual Inversion (with MSE + cosine + smooth-L_1 loss)	*txt2img*	SD V1.5	10	No. of TI vectors
CN: ControlNet	*img2img*	SD V1.4 SD V1.5 SDXL	1	Depth map Canny edge
CN-TI: ControlNet with TI	*txt2img+ img2img*	SD V1.5	1	Depth map
CN-IP: ControlNet with Inpainting	*txt2img+ img2img*	SD V1.5	1	Depth map Segmt mask

open-source.

3. Preliminaries

DreamBooth [33] is a form of latent diffusion model [32] that utilizes the latent space of denoising autoencoders for noise estimation and cross-attention layers to allow conditioning text to create text-to-image generators for high resolution image synthesis. It performs fine-tuning using (i) a few samples of an entity, (ii) text prompts containing a unique identifier and the class label associated with that entity, and (iii) regularization set for class/domain adaptation. DreamBooth fine-tunes the diffusion model while keeping the text and image encoders frozen.

Textual Inversion [13] trains personalized image generation model via a pre-trained text encoder while keeping the diffusion model frozen. It creates a new word embedding using a few example images of the new concept that the the model needs to learn. This technique updates text embeddings, such that the learned embeddings capture details corresponding to the example images.

ControlNet [43] is a neural network structure that learns conditional control via zero-initialized convolution layers that progressively updates for large pre-trained T2I diffusion models. It reuses the large-scale pre-trained layers of source models to build a deep and strong encoder to learn specific conditions in the form of alternate representations of the input signal, such as edge map, pose, etc. For an input image (latent encoding) z_0, image diffusion algorithms progressively add noise, $\varepsilon \sim \mathcal{N}(0,1)$ to produce a noisy image (latent) z_t, where t represents the number of times noise is added. At time step t, the diffusion algorithm uses the text prompt c_t and a feature encoding of task-specific conditioning input c_f to predict the noise via a parameterized de-noising network ε_θ using the objective function $\mathscr{L}_{CN}$, where, $\mathscr{L}_{CN} = \mathbb{E}_{z_0,t,c_t,c_f,\varepsilon} \| \varepsilon - \varepsilon_\theta(z_0,t,c_t,c_f) \|$ for finetuning diffusion models with ControlNet. ControlNet can be used as a stand-alone *img2img* module that uses a reference image and produces a conditional input such as Canny edge map from the input. In the *txt2img* variant, ControlNet uses the text to introduce diversity in the synthesized image controlled via *controlnetscale_guidance* parameter.

Diffusion-based Inpainting assigns values to the image pixels based on their proximity to the missing or damaged area represented by a binary mask, m via heat diffusion. The heat equation is applied to the missing/unknown values, which redistributes the intensity values to create a seamless patch. At each iteration, the area not covered by the mask is substituted with the original image latents plus the amount of noise for that step. This ensures the area under the mask is retained, $z_t^{known} = z_0$ and the diffusion model only changes the masked area, $z_t^{unknown}$. The objective function is $\mathscr{L}_{IP}$, where, $\mathscr{L}_{IP} = \mathbb{E}_{z_0,t,c_t,m,\varepsilon} \| \varepsilon - \varepsilon_\theta(z_t,c_t,m) \|$. Here, $z_{t-1} = m \odot z_t^{known} + (1-m) \odot z_t^{unknown}$. Several inpainting methods using diffusion models exist, such as SD-Edit [32], DiffEdit [9] and SmartBrush [40] that work on generic objects (DiffEdit, SmartBrush) and may not apply to face images while Collaborative Diffusion [16] was used to edit only age and beard in faces.

Existing GAN and Diffusion-based facial attribute editing and semantic inpainting methods either fail to successfully edit the desired attribute or degrade the identity-specific details. We propose to address this issue through unified local and global editing frameworks.

4. Proposed Method

We formulate the problem of identity-preserving facial attribute editing at two levels: *global* and *local*. Global editing affects the entire image, and therefore, we need a method that can preserve high-level details of the individual while generating diverse outputs consistent with the text prompt. We use the method proposed in [2] which uses a DreamBooth-based regularization approach that fine-tunes the diffusion model using contrastive loss for age editing. We adapt this approach by modifying the regularization set with few exemplars (30 images) belonging to each attribute resulting in a total of 780 images. The overall idea is to augment the prior knowledge of a pre-trained model about the generic *person* class while learning the distinctive characteristics of each attribute in a contrastive setting. We use the Stable Diffusion V1.5 model as the backbone and use the following objective function.

$$\mathcal{L}_{DB-prop} = \mathbb{E}_{z_0,t,c_t,\varepsilon}[\|\varepsilon - \varepsilon_\theta(z_0,t,c_t)\| + \\ \lambda_p\|\varepsilon - \varepsilon_\theta(z_0,t,c_t,c_{prior})\| + \lambda_s\mathcal{S}(\varepsilon,\varepsilon_\theta(z_0,t,c_t))]. \tag{1}$$

In Eqn. 1, the first term, denotes the squared error between the input images and the generated images; the second term denotes the prior-preservation loss associated with the class *person*; the third term $\mathcal{S}(\cdot,\cdot)$ denotes the temperature-scaled cross entropy loss in contrastive learning between the latent representation of the noise-free inputs (z_0) and the de-noised outputs (z_t) with a weight term λ_s and a temperature value = 0.5. Refer to [2] for details. We further augment our approach by using textual inversion-based regularization (different from DreamBooth) with cosine embedding loss and smooth-L_1 loss.

$$v_* = \arg_v\min\mathcal{L}_{TI} = \arg_v\min\mathbb{E}_{z_0,t,c_t,\varepsilon}[\|\varepsilon - \varepsilon_\theta(z_0,t,c_t)\| + \\ \lambda_{sl}\mathcal{S}_{L1}(\varepsilon,\varepsilon_\theta(z_0,t,c_t)) + \lambda_c\mathcal{C}(\varepsilon,\varepsilon_\theta(z_0,t,c_t))]. \tag{2}$$

In Eqn. 2, $\mathcal{C}(\cdot,\cdot)$ denotes the cosine embedding loss and $\mathcal{S}_{L1}(\cdot,\cdot)$ denotes the smooth-L_1 loss is governed by a quadratic function when the absolute element-wise error is close to zero, otherwise, it converges to L_1 loss. The optimization uses the same training paradigm as that of a latent diffusion model while ensuring the text encoder learns a special embedding, v_* pertaining to the identity of a specific concept/person.

Local editing, on the other hand, needs conditional generation such that only a small portion of the image undergoes change while preserving the rest of the input. This allows zero-shot editing using a single image, *i.e.*, regularization free approach and perform multi-attribute editing by strategically combining conditional masks. As local editing involves perturbing a small area compared to the entire image, we need to ensure smooth transitioning. This needs a conditioning input that captures fine-grained facial details in the form of an edge map or a depth map from the original input. So, we use a mask-guided inpainting approach with the mask as an auxiliary control. The method requires three inputs: (i) a binary semantic segmentation mask, $m(x,y)$ that denotes the pixels that need to be edited as foreground with value 1 and the remaining background pixels as 0.

$$m(x,y) = \begin{cases} 1 & \text{if } (x,y) \in foreground \\ 0 & \text{else if } (x,y) \in background \end{cases} \tag{3}$$

(ii) a depth map or depth image, $d(x,y)$ is a representation of the distance or depth information for each pixel in the original scene. The depth image is particularly suitable for capturing intricate facial details. Alternatively, a 2-D edge map can be used but we observe better facial detail preservation with depth maps. We combine $\mathcal{L}_{CN}$ and $\mathcal{L}_{IP}$ to have the following objective.

$$\mathcal{L}_{CN-IP} = \mathbb{E}_{z_0,z_t,t,c_t,d_f,m,\varepsilon}\|\varepsilon - \varepsilon_\theta(z_0,t,d_f,z_t,c_t,m)\| \tag{4}$$

In Eqn. 4, d_f denotes the feature encoding of the conditional input, *i.e.*, the depth map of the original face image, m denotes the mask supplied as auxiliary input for local attribute control, the editing prompt c_t that should correspond to the mask, *i.e.*, we do not ask the diffusion model to edit eyebrows while using a mask indicating the hair region. In such cases, the output image is a reconstruction of the original input. Similarly, for empty text prompts, *i.e.*, $c_t = \emptyset$, the output corresponds to a slight variant of the input.

| No attrib | Big nose | Blond hair | Brown hair | Dub chin | Eyeglasses | Hat | Mustache | Bald |

| Bangs | Big lips | Black hair | Necktie | Eyebrows | No beard | Mo open | Old | Young |

| Male | Female | Smiling | Angry | Surprise | Disgust | Sad | Fear | Neutral |

Figure 2. Outputs on LFW dataset using the baseline method **DB-base**. 'No attrib': No attribute editing; 'Dub chin': Double chin; 'Eyebrows': Bushy eyebrows and 'Mo open': Mouth slightly open.

5. Experiments

We conduct experiments using 300 subjects, 100 subjects each from the CelebA [27], CelebAMaskHQ [22] and LFW [15] datasets. CelebA and LFW subjects were selected such that each subject has at least 10 images. We selected the first 100 subjects from the CelebAMaskHQ dataset for our experiments which comes with semantic segmentation masks. We ablate two proposed methods, DreamBooth-proposed with contrastive loss (DB-prop.) for global editing and ControlNet with In-painting (CN-IP) for local editing. DB-prop. uses 780 images (30×26) for regularization set that contains image-caption pairs pertaining to 26 attributes. We select subject disjoint CelebA images for facial and demographic attributes and images from AffectNet [28] dataset for expression-based images in the regularization set. We use prompts as "photo of a ⟨ rare identifier ⟩ ⟨ person ⟩ with/wearing ⟨ attribute ⟩". Here, the rare identifier corresponds to a unique token that binds a specific identity to the class (we follow the procedure described in [2]). For CN-IP, we use the SDV1.5 with Control Net and the Dense Prediction Transformer (DPT)-Hybrid model (MiDaS 3.0) [31] for depth maps and Canny for edge maps.

Biometric evaluation: We use the ArcFace [34] and AdaFace [20] pre-trained models with RetinaFace [11] detector for biometric matching between original images (gallery set contains 3 images per subject) and the generated images. *No attribute* operation assesses reconstruction of original image in the absence of attribute edit. We observed initially extremely high values of FNMR on LFW dataset. So, we filtered out the poor quality images lowering the gallery set to 224 from 300. We use 200 images (2 images per subject) for gallery set of CelebAMaskHQ.

Attribute editing evaluation: Recent work on attribute prediction uses ChatGPT [10] that achieves an avg. $\sim 77\%$ accuracy on MAAD dataset and outperforms 4 out of 47 attributes compared to custom ResNet-50 models trained on specific attributes. We use a pre-trained LLaVA-v1.5-7b [24] that achieves SoTA on 11 benchmarks and surpasses methods like Qwen-VL-Chat that use billion-scale data. We use LLaVA in the open-ended visual-question answering mode for attribute prediction. We prompt the LLaVA model pertaining to an attribute-edited image and if the response matches with the desired attribute edit, we consider that as a success. For example, for the prompt *photo of a person with black hair*, we prompt the LLaVA model using the image generated by the diffusion model with the following question: *Does the person have black hair?* If the LLaVA model generates response as *Yes*, we consider that as a successful attribute editing operation. We compute the proportion over all subjects and report the results.

Baselines: We use DreamBooth-base (DB-base) with Stable Diffusion (SD) V1.5 as the baseline for regularization-based method for global editing. We use the source code provided by auhtors of InstantID [39] and BLIPDiffusion [23] as regularization-free baselines for local editing. We use both DreamBooth and textual inversion (TI) for subject and attribute-specific fine-tuning.
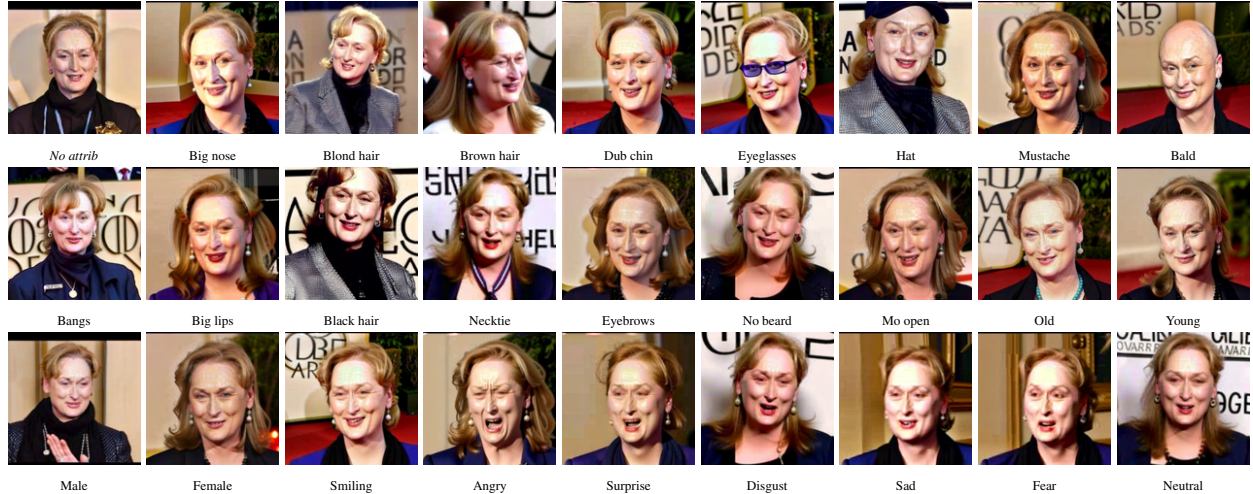
Figure 3. Outputs on LFW dataset using our global attribute editing method **DB-prop**. Note *Bald*, *Angry*, *Male*, *Female* edited images maintain original identity in contrast to DB-base in Fig. 2.
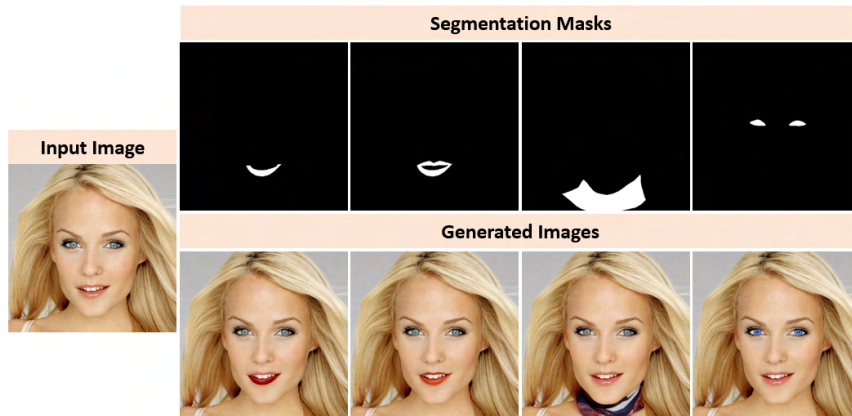

Figure 4. Fine-grained *local* attribute editing by our CN-IP method. The first image corresponds to editing only the *lower lip*. The second image edits *both lips* to orange color. We achieve this by adding the lower and upper lip masks and then performing the editing operation. The third image corresponds to adding necktie using the *neck* mask and the fourth image corresponds to editing the eye color to blue by adding both left and right eye masks.

6. Results

Qualitative Analysis. We present examples of global attribute editing by DB-base and DB-prop. in Figs. 2 and 3, respectively. DB-prop. edits attributes while preserving perceptual quality with reference to the original individual. We present examples of fine-grained local attribute editing using the proposed CN-IP method in Fig. 4. Note CN-IP performs nuanced attribute editing guided by segmentation mask and depth map that preserves the overall facial details.

Quantitative Analysis. We present biometric matching results using ArcFace and AdaFace and attribute prediction results using LLaVA on both global and local frameworks.

Biometric evaluation—*Global:* We present the results for *global* attribute editing on the **CelebA** dataset in Tab. 2 and on the **LFW** dataset in Tab. 3. We highlight those results where the degradation in FNMR is $\geq 10\%$ with ArcFace and $\geq 5\%$ with AdaFace matcher. We used these values because we observe, overall, AdaFace outperforms ArcFace on 'Original' images and 'No attribute' images. We observe that several attribute editing operations significantly degrade biometric recognition (see the red-colored cells). Our findings support that digital attribute editing operations can circumvent FR systems. DB-base reports significant drop in performance on LFW dataset using ArcFace by up to 87% FNMR@FMR=0.01% on *black hair*. The proposed global editing method, DB-prop., can handle attribute editing while successfully preserving the identity (see the green-colored cells). Note we report the results on all 26 attributes + No attribute for analysis.
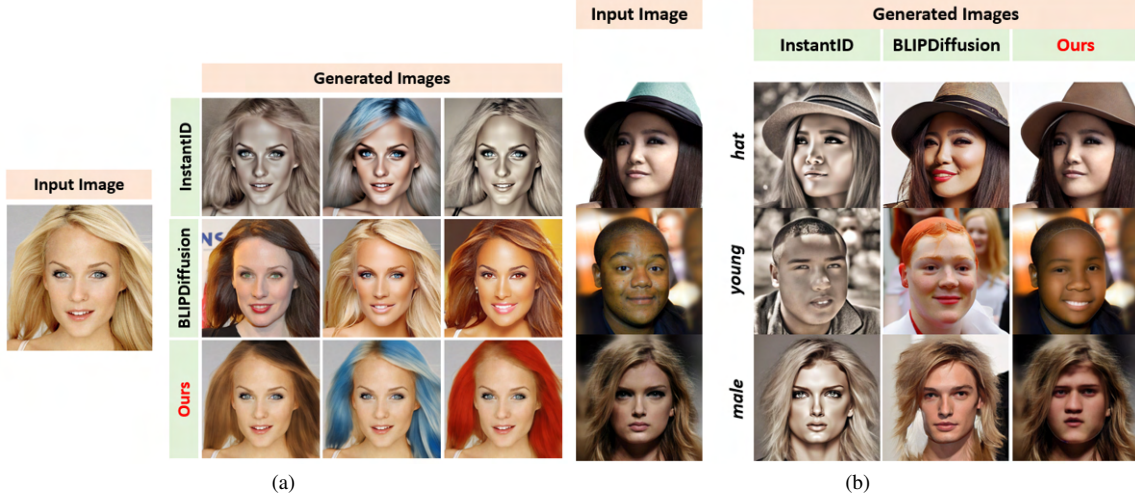
Figure 5. Baselines (InstantID and BLIPDiffusion) vs. Our Proposed method (CN-IP). In (a), we show examples of generated images for *unseen* attributes. Note BLIPDiffusion requires a reference image so in absence of a reference image of a person with blue hair, it defaults to the reconstruction of the original image. InstantID is unable to successfully perform editing for rare unseen attributes. In (b), we show examples pertaining to three different types of attributes: *hat*, *young* and *male*.
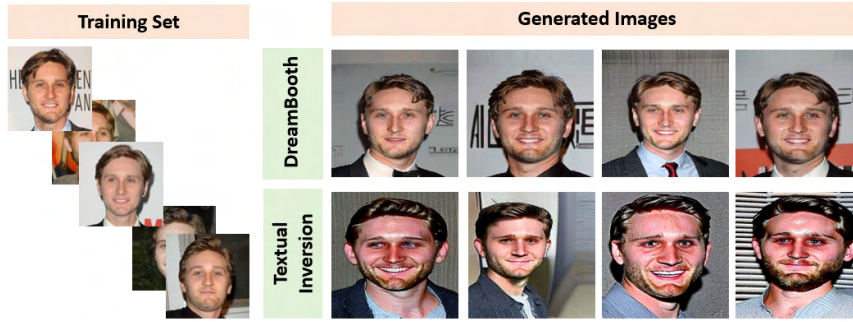

Figure 6. DreamBooth vs. Textual inversion for subject-specific fine-tuning. We use the same training set of 10 images in both fine-tuning methods. In DreamBooth, we use the contrastive loss; in textual inversion we use cosine embedding loss with smooth-L_1 loss and set the number of vectors to learn the concept=5. Our global editing method, DreamBooth (DB-prop.) results in more visually realistic subject renditions compared to Textual Inversion.
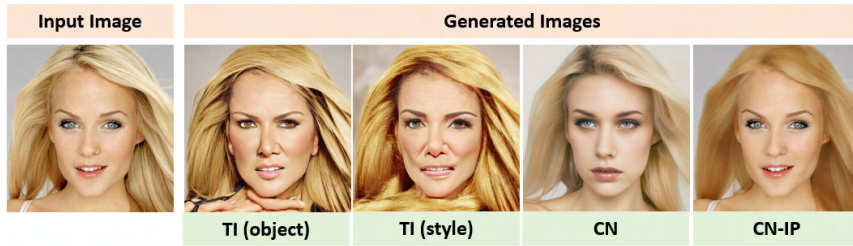

Figure 7. ControlNet vs. Textual inversion for attribute-specific fine-tuning. We use a set of 10 images of people with blond hair to learn the *blond* attribute using textual inversion. TI allows to learn a concept as object as well as style. After learning, we can re-use the embeddings on any individual. Here, we are showing reconstruction operation using the input image who also has blond hair. Both TI and CN fail to preserve the identity but our local editing method, CN-IP, faithfully preserves the details.

Biometric evaluation—*Local:* We present the results for *local* attribute editing, *i.e.*, our CN-IP model on the **CelebA-MaskHQ** dataset in Tab. 4. One of the baselines BLIPDiffusion reports a significant drop in performance by up to 96% FNMR@FMR=0.01% with ArcFace and by up to 94% FNMR@FMR=0.01% with AdaFace on *mouth slightly open*. InstantID on the other hand typically returns the original image with slight variations resulting in better identity matching but cannot process 30% of the images due to inability to detect faces even in high-resolution images (1024 × 1024) of CelebAMaskHQ

Table 2. FNMR(↓)@FMR(%)=0.01/0.1 DreamBooth (base) vs. DreamBooth (proposed) on **CelebA**. Red-colored cells denote degradation in FNMR of attribute edited images using the baseline (DB-base) by ≥ 10% with ArcFace and ≥ 5% with AdaFace compared to original performance. Green-colored cells denote that our method (DB-proposed) shows consistent improvement in FNMR of attribute edited images.

	ArcFace		AdaFace	
Matcher	DB-base	DB-prop.	DB-base	DB-prop.
Attribute	@FMR	@FMR	@FMR	@FMR
Type	0.01/0.1	0.01/0.1	0.01/0.1	0.01/0.1
Original	0.33/0.09	0.33/0.09	0.08/0.07	0.08/0.07
Facial semantic attributes				
No attribute	0.40/0.05	0.33/0.11	0.09/0.05	0.09/0.08
bald	0.43/0.14	0.38/0.13	0.15/0.08	0.09/0.08
bangs	0.41/0.13	0.45/0.14	0.09/0.06	0.09/0.07
big lips	0.28/0.07	0.44/0.12	0.05/0.04	0.09/0.07
big nose	0.46/0.17	0.40/0.13	0.11/0.05	0.09/0.07
black hair	0.51/0.18	0.38/0.11	0.13/0.09	0.09/0.07
blond hair	0.32/0.11	0.40/0.13	0.07/0.04	0.09/0.06
brown hair	0.39/0.11	0.40/0.12	0.08/0.04	0.08/0.06
bushy eyebrows	0.35/0.12	0.38/0.12	0.07/0.04	0.09/0.07
double chin	0.46/0.17	0.40/0.11	0.11/0.07	0.08/0.06
eyeglasses	0.30/0.04	0.50/0.16	0.08/0.05	0.08/0.06
hat	0.39/0.11	0.41/0.14	0.10/0.05	0.09/0.07
mustache	0.32/0.08	0.37/0.12	0.06/0.04	0.09/0.07
necktie	0.47/0.14	0.35/0.12	0.09/0.05	0.08/0.06
no beard	0.34/0.10	0.38/0.12	0.08/0.05	0.09/0.06
mouth slightly open	0.28/0.08	0.38/0.13	0.09/0.05	0.08/0.05
Demographic attributes				
female	0.55/0.22	0.44/0.16	0.17/0.09	0.10/0.07
male	0.41/0.14	0.38/0.14	0.12/0.07	0.08/0.05
old	0.48/0.18	0.39/0.13	0.21/0.13	0.08/0.06
young	0.41/0.12	0.37/0.12	0.10/0.06	0.09/0.06
Expression attributes				
anger	0.62/0.24	0.40/0.13	0.21/0.11	0.08/0.06
happy	0.28/0.06	0.37/0.13	0.07/0.05	0.09/0.06
fear	0.31/0.07	0.36/0.13	0.08/0.05	0.08/0.07
disgust	0.31/0.10	0.39/0.14	0.10/0.05	0.08/0.07
sad	0.36/0.11	0.40/0.12	0.09/0.05	0.08/0.06
surprise	0.41/0.18	0.39/0.13	0.09/0.05	0.09/0.07
neutral	0.31/0.08	0.37/0.11	0.07/0.04	0.08/0.07

dataset.

Key Findings: 1. DB-prop. reduces FNMR in all cases on CelebA within an avg. ∼ 5.4% of the original performance using ArcFace and within an avg. ∼ 1.3% of the original performance using AdaFace. On LFW, DB-prop. outperforms DB-base in a majority of cases, within an avg. ∼ 14% of the original performance using ArcFace and within an avg. ∼ 3.2% of the original performance using AdaFace.

2. CN-IP maintains an avg. FNMR within ∼ 6.8% of the original performance using ArcFace and within ∼ 3.1% of the original performance using AdaFace.

3. The following attribute editing operations impair biometric matching across CelebA and LFW datasets with both ArcFace and AdaFace matchers : *black hair*, *bald*, *male*, *female*, *old* and *anger*.

Attribute prediction We present results of attribute prediction using the LLaVA model for both global (DB-prop.) and local attribute editing (CN-IP) methods in Tab. 5. We observe that BLIPDiffusion outperforms our global and local attribute editing methods in 9 out of 18 attributes. Note that on both *young* and *old* attributes our CN-IP model achieves 98% (close to 100% of BLIP). BLIP requires a reference image as the style image, so for each attribute we provided an example image. Both

Table 3. FNMR(↓)@FMR(%)=0.01/0.1 DreamBooth (base) vs. DreamBooth (proposed) on **LFW**. Red-colored cells denote degradation in FNMR of attribute edited images using the baseline (DB-base) by ≥ 10% with ArcFace and ≥ 5% with AdaFace compared to original performance. Green-colored cells denote that our method (DB-proposed) shows improvement in FNMR of attribute edited images in a majority of cases.

	ArcFace		AdaFace	
Matcher	DB-base	DB-prop.	DB-base	DB-prop.
Attribute	@FMR	@FMR	@FMR	@FMR
Type	0.01/0.1	0.01/0.1	0.01/0.1	0.01/0.1
Original	0.10/0.01	0.10/0.01	0.05/0.04	0.05/0.04
Facial semantic attributes				
No attribute	0.24/0.05	0.11/0.01	0.07/0.05	0.07/0.05
bald	0.49/0.19	0.18/0.04	0.16/0.07	0.05/0.04
bangs	0.14/0.03	0.17/0.04	0.05/0.04	0.06/0.03
big lips	0.14/0.02	0.30/0.04	0.05/0.05	0.09/0.05
big nose	0.28/0.05	0.42/0.04	0.07/0.05	0.12/0.09
black hair	0.87/0.06	0.19/0.03	0.16/0.11	0.08/0.05
blond hair	0.31/0.03	0.34/0.03	0.08/0.06	0.08/0.06
brown hair	0.26/0.04	0.24/0.03	0.06/0.05	0.09/0.07
bushy eyebrows	0.23/0.04	0.28/0.04	0.08/0.07	0.10/0.07
double chin	0.67/0.05	0.33/0.03	0.09/0.07	0.09/0.06
eyeglasses	0.42/0.12	0.25/0.04	0.06/0.05	0.06/0.05
hat	0.52/0.07	0.22/0.03	0.08/0.05	0.06/0.05
mustache	0.71/0.08	0.21/0.03	0.11/0.05	0.07/0.05
necktie	0.37/0.07	0.16/0.03	0.13/0.10	0.06/0.04
no beard	0.45/0.07	0.29/0.05	0.15/0.11	0.08/0.06
mouth slightly open	0.32/0.06	0.44/0.04	0.06/0.05	0.10/0.07
Demographic attributes				
female	0.59/0.27	0.18/0.04	0.25/0.14	0.08/0.05
male	0.83/0.20	0.10/0.02	0.17/0.10	0.06/0.05
old	0.43/0.16	0.13/0.03	0.17/0.09	0.04/0.03
young	0.40/0.12	0.12/0.02	0.08/0.05	0.06/0.04
Expression attributes				
anger	0.60/0.20	0.21/0.03	0.18/0.11	0.06/0.04
happy	0.16/0.01	0.26/0.02	0.05/0.05	0.09/0.05
fear	0.21/0.04	0.19/0.03	0.05/0.04	0.07/0.05
disgust	0.19/0.03	0.29/0.04	0.06/0.05	0.07/0.05
sad	0.30/0.06	0.18/0.03	0.08/0.06	0.07/0.05
surprise	0.29/0.07	0.25/0.03	0.11/0.08	0.07/0.06
neutral	0.19/0.02	0.19/0.02	0.05/0.05	0.06/0.04

InstantID and CN-IP methods rely on text prompts for attribute editing.

Key Findings: 1. Attribute editing works well if a reference image is used for style transfer as followed in BLIPDiffusion implementation [4]. However, certain attributes such as *bald* and *black hair* are difficult to induce

2. InstantID performs the worst in terms of attribute editing even with lower *ip_adapter* scale for better prompt control [17].

3. Our global and local editing methods together outperform both InstantID and BLIPDiffusion showing the need for considering attribute editing at two distinct levels. BLIPDiffusion cannot work on rare and unseen attributes due to absence of reference image demonstrating the robustness of our methods.

Comparison with existing methods. We compare with 3 baselines: DB-base for global and BLIPDiffusion and InstantID for local editing. For global editing, see Fig. 2 for qualitative comparison, Tabs. 2 and 3 for biometric evaluation comparison and Tab. 5 for attribute prediction comparison. For local editing, see Fig. 5 for qualitative comparison, Tab. 4 for biometric evaluation comparison and Tab. 5 for attribute prediction comparison.

Key Findings: 1. For global editing, DB-prop. significantly outperforms DB-base in terms of biometric evaluation by 22% (ArcFace-CelebA), by 13% (AdaFace-CelebA), by 73% (ArcFace-LFW) and by 17% (AdaFace-LFW).

Table 4. FNMR(↓)@FMR(%)=0.01/0.1 CN-IP (ours) vs. BLIPDiffusion and InstantID (baselines) for local attribute editing on **CelebA-MaskHQ** with ArcFace and AdaFace. CN-IP outperforms both baselines in a majority of cases by a significant margin.

	ArcFace			**AdaFace**		
Matcher	BLIP	InstantID	CN-IP	BLIP	InstantID	CN-IP
Attribute Type	@FMR 0.01/0.1	@FMR 0.01/0.1	@FMR 0.01/0.1	@FMR 0.01/0.1	@FMR 0.01/0.1	@FMR 0.01/0.1
Original	0.15/0.02	0.15/0.02	0.15/0.02	0.04/0.04	0.04/0.04	0.04/0.04
bald	0.9/0.64	0.27/0.13	0.21/0.08	0.77/0.42	0.09/0.06	0.07/0.06
blond hair	0.83/0.5	0.26/0.09	0.20/0.07	0.61/0.26	0.08/0.05	0.07/0.06
black hair	0.78/0.49	0.25/0.11	0.21/0.08	0.64/0.26	0.06/0.05	0.07/0.06
brown hair	0.81/0.51	0.23/0.10	0.20/0.07	0.64/0.30	0.06/0.05	0.07/0.06
bangs	0.88/0.5	0.21/0.11	0.22/0.08	0.68/0.32	0.08/0.07	0.07/0.06
bushy eyebrows	0.86/0.56	0.24/0.11	0.21/0.08	0.72/0.41	0.08/0.08	0.07/0.06
big nose	0.89/0.62	0.24/0.10	0.29/0.08	0.79/0.47	0.07/0.05	0.07/0.06
big lips	0.82/0.5	0.24/0.13	0.19/0.07	0.67/0.29	0.08/0.06	0.07/0.06
necktie	0.86/0.56	0.24/0.13	0.22/0.06	0.72/0.44	0.08/0.07	0.07/0.06
mouth slightly open	0.85/0.57	0.23/0.13	0.23/0.07	0.63/0.34	0.07/0.06	0.08/0.07
hat	0.96/0.76	0.25/0.13	0.22/0.06	0.94/0.78	0.08/0.06	0.07/0.06

Table 5. Attribute prediction using LLaVA on **CelebAMaskHQ**.

Attribute	Instant ID	BLIP Diffusion	DB-prop. (Ours)	CN-IP (Ours)
bald	7.25	50	27	15.31
black hair	15.94	69	41	50
blond hair	97.10	100	67	90.82
brown hair	55.07	99	74	92.86
bangs	85.51	70	63	61.22
big nose	65.22	100	100	80.81
big lips	100	100	100	100
bushy eyebrows	98.55	100	100	100
slightly open mouth	100	100	100	100
young	94.20	100	89	98
old	49.28	100	25	98
male	15.94	91	35	64
female	95.65	92	88	92
neutral	72.46	73	63	71
hat	30.43	100	72	68.37
necktie	50.72	60	83	63
eyeglasses	5.80	51	84	100

2. For local editing, CN-IP method outperforms BLIPDiffusion by 74% (ArcFace) and by 87% (AdaFace); outperforms InstantID by 6% (ArcFace) and by 2% (AdaFace).

3. In terms of attribute editing, BLIPDiffusion outperforms proposed method by 28% on *hat* while our methods outperform BLIPDiffusion by 49% and InstantID by 94.2% on *eyeglasses*.

4. In [1], the authors report ArcFace performance after using AttGAN [14] for attribute editing on 100 subjects from LFW (subjects may not be identical to our set).Nonetheless, we observe that DB-prop significantly reduces the FNMR on all attribute editing operations specially on *eyeglasses* (by 74%) and *bangs* (by 69%).

Additional evaluation. Firstly, we compare subject-specific fine-tuning and attribute-specific fine-tuning using textual inversion. See Figs. 6 and 7. We use variable length of embedding vectors {1,2,5} for TI to learn identity and attribute as style/object. However, we observe that DreamBooth produces more realistic looking subject renditions while CN-IP produces more realistic attribute editing renditions compared to TI. Secondly, we compare two VQA models BLIP (using

pre-trained BERT encoder) and LLaVA on $\sim 2,500$ images from CelebAMaskHQ original dataset which comes with ground truth attribute annotations. We observe that BLIP achieves only 48.5% success on attribute prediction (fails completely on eyeglass detection) while LLaVA model achieves 59.3%. Thirdly, we use different model versions of SDV1.5, SDV2.1 and SDXL. Although SDXL produces highly realistic images but when used with ControlNet, the outputs deviate from the original identity. So, SDV1.5 is used in DB-prop. and CN-IP.

Limitations. We observe that the regularization-based method works better on global attributes such as age and facial expression editing while the regularization-free method works better on fine-grained local attribute editing. For expression editing using the CN-IP model, we provide the face mask as conditional input which results in the inpainting model to hallucinate facial details producing random faces.

7. Conclusion

Facial attribute editing may often be presented as innocuous post-processing, but an adversary could leverage editing tools to increase false match or non-match. We study SoTA diffusion models and use them for modifying a large suite of semantic, demographic and expression attributes on three datasets (CelebA, CelebAMaskHQ and LFW). We observe several attribute editing operations (altering sexual cues, age and hair style) significantly degrade biometric matching. Therefore, we propose two mitigation techniques specifically designed to handle both global and local attribute editing tasks while preserving identity. We use DreamBooth with attribute-based regularization set in a contrastive setting for global editing and a ControlNet model with Stable Diffusion guided by segmentation masks and depth maps for local attribute editing. The proposed methods can faithfully preserve biometric fidelity and perceptual quality and outperform both GAN (AttGAN) and diffusion (DB-base, InstantID and BLIPDiffusion) baselines with a significant margin. We further use LLaVA as a VQA tool for attribute prediction and observe promising results on our methods.

References

[1] S. Banerjee, A. Aggarwal, and A. Ross. Can gan-induced attribute manipulations impact face recognition? *International Conference in Pattern Recognition Workshops (UMDBB)*, 2022. 1, 2, 10

[2] S. Banerjee, G. Mittal, A. Joshi, C. Hegde, and N. Memon. Identity-preserving aging of face images via latent diffusion models. *International Joint Conference in Biometrics (IJCB)*, 2023. 2, 4, 5, 13, 15

[3] A. Bharati, R. Singh, M. Vatsa, and K. W. Bowyer. Detecting facial retouching using supervised deep learning. *IEEE Transactions on Information Forensics and Security*, 11(9):1903–1913, 2016. 1, 2

[4] BLIPDiffusion Implementation. `https://github.com/salesforce/LAVIS/tree/main/projects/blip-diffusion`. [Online accessed: 3rd Feb, 2024]. 9

[5] H.-J. Chen, K.-M. Hui, S.-Y. Wang, L.-W. Tsao, H.-H. Shuai, and W.-H. Cheng. BeautyGlow: On-Demand Makeup Transfer Framework With Reversible Generative Network. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 10034–10042, 2019. 1, 2

[6] X. Chen and S. Lathuilière. Face aging via diffusion-based editing. *BMVC*, 2023. 2

[7] J. Choi, J. Lee, C. Shin, S. Kim, H. Kim, and S. Yoon. Perception prioritized training of diffusion models. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1

[8] C.-Y. Chuang, J. Varun, Y. Li, A. Torralba, and S. Jegelka. Debiasing vision-language models via biased prompts. *arXiv preprint 2302.00070*, 2023. 2

[9] G. Couairon, J. Verbeek, H. Schwenk, and M. Cord. Diffedit: Diffusion-based semantic image editing with mask guidance. In *International Conference on Learning Representations ICLR*, 2023. 1, 3

[10] I. Deandres-Tame, R. Tolosana, R. Vera-Rodriguez, A. Morales, J. Fierrez, and J. Ortega-Garcia. How good is chatgpt at face biometrics? a first look into recognition, soft biometrics, and explainability. *IEEE Access*, 2024. 5

[11] J. Deng, J. Guo, E. Ververas, I. Kotsia, and S. Zafeiriou. Retinaface: Single-shot multi-level face localisation in the wild. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 5

[12] J. Deng, J. Guo, N. Xue, and S. Zafeiriou. ArcFace: Additive Angular Margin Loss for Deep Face Recognition. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4685–4694, 2019. 1

[13] R. Gal, Y. Alaluf, Y. Atzmon, O. Patashnik, A. H. Bermano, G. Chechik, and D. Cohen-Or. An image is worth one word: Personalizing text-to-image generation using textual inversion. 2023. 3

[14] Z. He, W. Zuo, M. Kan, S. Shan, and X. Chen. AttGAN: Facial Attribute Editing by Only Changing What You Want. *IEEE Transactions on Image Processing*, 28(11):5464–5478, 2019. 2, 10

[15] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: A database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, October 2007. 5

[16] Z. Huang, K. C. Chan, Y. Jiang, and Z. Liu. Collaborative diffusion for multi-modal face generation and editing. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2023. 3

[17] InstantID Implementation. `https://github.com/InstantID/InstantID`. [Online accessed: 3rd Feb, 2024]. 9

[18] Y. Jiang, Z. Huang, X. Pan, C. C. Loy, and Z. Liu. Talk-to-edit: Fine-grained facial editing via dialog. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 13799–13808, 2021. 1, 2, 15, 17

[19] A. Joshi, A. Mukherjee, S. Sarkar, and C. Hegde. Semantic adversarial attacks: Parametric transformations that fool deep classifiers. In *Proceedings of the IEEE/CVF international conference on computer vision*, pages 4773–4783, 2019. 1

[20] M. Kim, A. K. Jain, and X. Liu. Adaface: Quality adaptive margin for face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, 2022. 1, 5

[21] J. Kwak, D. K. Han, and H. Ko. CAFE-GAN: Arbitrary Face Attribute Editing with Complementary Attention Feature. In *European Conference on Computer Vision (ECCV)*, 2020. 1, 2

[22] C.-H. Lee, Z. Liu, L. Wu, and P. Luo. Maskgan: Towards diverse and interactive facial image manipulation. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2020. 2, 5

[23] D. Li, J. Li, and S. C. H. Hoi. BLIP-Diffusion: Pre-trained Subject Representation for Controllable Text-to-Image Generation and Editing. *NeurIPS*, 2023. 2, 5

[24] H. Liu, C. Li, Q. Wu, and Y. J. Lee. Visual instruction tuning. 2023. 5

[25] M. Liu, Y. Ding, M. Xia, X. Liu, E. Ding, W. Zuo, and S. Wen. STGAN: A Unified Selective Transfer Network for Arbitrary Image Attribute Editing. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 3668–3677, 2019. 1, 2

[26] S. Liu, D. Li, T. Cao, Y. Sun, Y. Hu, and J. Ji. GAN-Based Face Attribute Editing. *IEEE Access*, 8:34854–34867, 2020. 1, 2

[27] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep Learning Face Attributes in the Wild. In *Proceedings of International Conference on Computer Vision*, December 2015. 5

[28] A. Mollahosseini, B. Hasani, and M. H. Mahoor. AffectNet: A Database for Facial Expression, Valence, and Arousal Computing in the Wild. *IEEE Trans. Affect. Comput.*, 10(1):18–31, jan 2019. 5

[29] M. Ngan, P. Grother, and A. Hom. Face Analysis Technology Evaluation (FATE) Part 10: Performance of Passive, Software-Based Presentation Attack Detection (PAD) Algorithms. *NIST Internal Report 8491* `https://nvlpubs.nist.gov/nistpubs/ir/2023/NIST.IR.8491.pdf`, 2023. 1

[30] M. V. Perera and V. M. Patel. Analyzing bias in diffusion-based face generation models. *International Joint Conference in Biometrics (IJCB)*, 2023. 2

[31] R. Ranftl, A. Bochkovskiy, and V. Koltun. Vision transformers for dense prediction. *CoRR*, abs/2103.13413, 2021. 5

[32] R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, June 2022. 3

[33] N. Ruiz, Y. Li, V. Jampani, Y. Pritch, M. Rubinstein, and K. Aberman. DreamBooth: Fine Tuning Text-to-image Diffusion Models for Subject-Driven Generation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2023. 2, 3

[34] S. I. Serengil and A. Ozpinar. Lightface: A hybrid deep face recognition framework. In *Innovations in Intelligent Systems and Applications Conference (ASYU)*, pages 23–27, 2020. 5

[35] S. I. Serengil and A. Ozpinar. Hyperextended lightface: A facial attribute analysis framework. In *International Conference on Engineering and Emerging Technologies (ICEET)*, pages 1–4, 2021. 13

[36] R. Sun, C. Huang, H. Zhu, and L. Ma. Mask-aware photorealistic facial attribute manipulation. In *Computational Visual Media*, volume 7, pages 363–374, 2021. 1, 2

[37] X. Tang, Z. Wang, W. Luo, and S. Gao. Face aging with identity-preserved conditional generative adversarial networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 7939–7947, 2018. 2

[38] P. Terhörst, J. N. Kolf, M. Huber, F. Kirchbuchner, N. Damer, A. M. Moreno, J. Fierrez, and A. Kuijper. A Comprehensive Study on Face Recognition Biases Beyond Demographics. *IEEE Transactions on Technology and Society*, 3(1):16–30, 2022. 1

[39] Q. Wang, X. Bai, H. Wang, Z. Qin, and A. Chen. Instantid: Zero-shot identity-preserving generation in seconds. *arXiv preprint arXiv:2401.07519*, 2024. 2, 5

[40] S. Xie, Z. Zhang, Z. Lin, T. Hinz, and K. Zhang. Smartbrush: Text and shape guided object inpainting with diffusion model. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 22428–22437, 2023. 3

[41] H. Yang, D. Huang, Y. Wang, and A. K. Jain. Learning Face Age Progression: A Pyramid Architecture of GANs. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 31–39, 2018. 1, 2

[42] D. Yue, Q. Guo, M. Ning, J. Cui, Y. Zhu, and L. Yuan. Chatface: Chat-guided real face editing via diffusion latent space manipulation, 2023. 2

[43] L. Zhang, A. Rao, and M. Agrawala. Adding conditional control to text-to-image diffusion models. In *IEEE International Conference on Computer Vision (ICCV)*, 2023. 3

A. Attribute prediction

A.1. Inference

We use LLaVA for attribute prediction in this work. The objective is to use LLaVA for attribute prediction and then assess the success of our editing method with respect to the editing prompts. For example, say our editing prompt states *photo of a person with mustache*, and we prompt the LLaVA model for the edited image with the question *Does the person have mustache?*. Now if the response of the LLaVA model is *Yes*, we consider that as a successful attribute editing operation, otherwise, we consider that as a failure case. We use the following questions to prompt the LLaVA model for all attributes considered in this work.

{*Is the person young?, Is the person old?, Is the person male?, Is the person female?, Is the person bald?, Does the person have black hair?, Does the person have blond hair?, Does the person have brown hair?, Does the person have bangs?, Does the person have no beard?, Does the person have mustache? Does the person have bushy eyebrows?, Does the person have double chin?, Does the person have slightly open mouth?, Does the person have big nose?, Does the person have big lips?, Is the person wearing eyeglasses?, Is the person wearing necktie?, Is the person wearing hat?, Is the person angry?, Is the person smiling?, Does the person have sad expression?, Does the person have fearful expression?, Does the person have neutral expression?, Does the person have disgusted expression?, Does the person have surprised expression?*}

Additionally, we use deepface analyze [35] for attribute prediction on DB-base model in Table 6 but deepface can only predict demographic attributes such as age, sex, ethnicity and emotional expression. This shows the need for LLMs as open-ended VQA models that can be used for large-scale attribute prediction.

A.2. Benchmark

We compared LLaVA with BLIP in visual-question answering mode by benchmarking it on 2,459 images from the original CelebAMaskHQ dataset in Table 7. We use the annotations provided by the authors as ground truth and prompt both models with the following set of questions. Note some of the questions are open-ended instead of yes/no prompts to gauge the robustness of the models. *e.g.*, *What is the emotional expression of the person in the photo?* vs. *Is the person happy?*

{*Does the person have Arched Eyebrows?, Is the person attractive?, Are there bags under eyes?, Is the person bald?, Are there hair bangs?, Does the person have big lips?, Does the person have big nose?, What is the hair color?, Are there bushy eyebrows?, Is the person chubby?, Does the person have double chin?, Is the person wearing glasses?, Is there a goatee?, Is the person wearing heavy makeup?, Does the person have high cheekbones?, What is the gender?, What is the age?, What is the ethnicity?, Is the mouth slightly open?, Does the person have mustache?, Does the person have beard?, Are the eyes narrow?, Is the face shape oval?, What is the emotional expression of the person in the photo?, What accessories are present in the photo?*}

B. Implementation Details

We use A100 GPU for training using DreamBooth and Textual Inversion. Each subject requires 10-15 mins. for training. We use the official implementation of DreamBooth-based facial age editing [2] for DB-prop. [1] We use HuggingFace for implementing textual inversion [2]. We use the ControlNetV1.1 with Stable Diffusion V1.5 in inpainting mode [3]. We use the

---

[1] https://github.com/sudban3089/ID-Preserving-Facial-Aging
[2] https://huggingface.co/docs/diffusers/en/using-diffusers/textual_inversion_inference
[3] https://huggingface.co/lllyasviel/control_v11p_sd15_inpaint

Table 6. Success rate of predicting edited facial attribute. Higher is better.

Attribute Category	Attribute Type	Dataset and SD model			
		CelebA SD v1.5	CelebA SD v2.0	LFW SD v1.5	LFW SD v2.0
Age	old	0.39	0.24	**0.66**	**0.78**
	young	**0.90**	**0.89**	0.48	0.33
Sex	male	**0.73**	**0.59**	**0.85**	**0.75**
	female	0.61	0.55	0.22	0.30
Expression	smiling	**0.95**	**0.93**	**1.00**	**0.98**
	angry	0.34	0.11	0.64	0.23
	disgust	0.00	0.01	0.00	0.00
	fear	0.06	0.08	0.01	0.13
	sad	0.18	0.23	0.36	0.20
	neutral	0.35	0.47	0.26	0.46

Table 7. Benchmark performance of BLIP vs. LLaVA on CelebAMaskHQ dataset. LLaVA outperforms BLIP in a majority of cases by a significant margin. Therefore, we use LLaVA in the attribute prediction experiments.

Attribute	BLIP	LLaVA
Arched eyebrows	**37.7**	36.7
Attractive	**59.1**	55.1
Bags under eyes	30.8	**52.7**
Bald	19.3	**96.7**
Bangs	20.5	**38.7**
Big lips	37.1	**48.4**
Big nose	34.2	**47.8**
Black hair	83.0	**84.7**
Blond hair	83.5	**87.1**
Brown hair	**65.7**	64.8
Bushy eyebrows	**20.2**	20.0
Chubby	64.0	**81.5**
Double chin	20.7	**46.5**
Eyeglasses	6.6	**94.5**
Goatee	39.8	**83.7**
Heavy makeup	56.1	**75.4**
High cheekbones	**47.5**	46.7
Mouth slightly open	45.6	**46.7**
Mustache	31.9	**78.5**
Narrow eyes	**15.8**	12.7
No beard	**79.6**	8.5
Oval face	**31.8**	20.6
Earrings	80.1	**83.9**
Hat	**95.2**	93.4
Lipstick	**45.3**	45.0
Necklace	**65.9**	59.8
Necktie	**92.1**	91.6

ArcFace matcher from the deepface library [4] and the AdaFace matcher from its original implementation [5]. We use the BLIP Diffusion from the HuggingFace [6]. We use LLaVA from the HuggingFace implementation [7]. We will release the codes for reproducibility.

C. Additional examples from CelebA

We present results of images generated using DB-base and DB-prop. on CelebA dataset in Figs. 8 and 9, respectively. We selected SDV2.0 model for DB-base and SDV1.5 model for DB-prop. in this example.

t-SNE analysis. We analyze the AdaFace embedding for the generated images particularly for the attributes that are affecting reliable face recognition. We use t-SNE [?] to reduce the 512-D feature vector to 3-D and present the plots for hair color (black hair, blonde hair and brown hair) for InstantID, BLIPDiffusion and CN-IP (ours) in Fig. 10. We observe that BLIPDiffusion forms well-defined clusters for different hair colors while in the case of Instant-ID and CN-IP, the embeddings are evenly distributed. Identity embeddings should not adhere to hair color variations, the tight clustering of embeddings from BLIPDiffusion indicates the contrary. As a result, we observe worse biometric performance from BLIPDiffusion compared to both InstantID and CN-IP.

---

[4] https://github.com/serengil/deepface
[5] https://github.com/mk-minchul/AdaFace
[6] https://huggingface.co/docs/diffusers/en/api/pipelines/blip_diffusion
[7] https://huggingface.co/docs/transformers/main/en/model_doc/llava
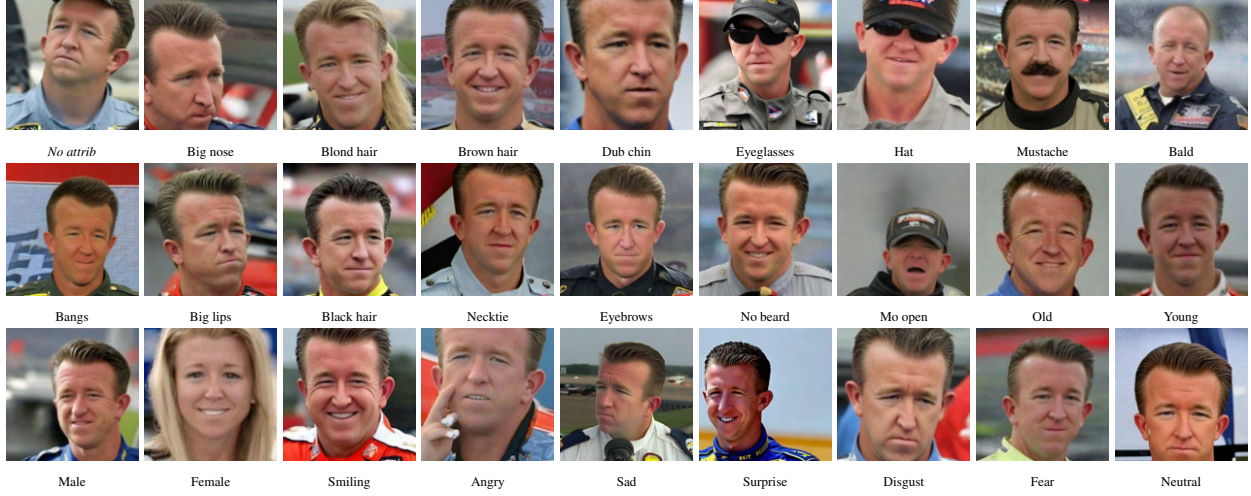
Figure 8. Example from CelebA dataset attributed edited using Stable Diffusion v2.0 model. 'No attrib': No attribute editing; 'Dub chin': Double chin; 'Eyebrows': Bushy eyebrows; 'Mo open': Mouth slightly open.


Figure 9. Example from CelebA dataset attributed edited using Stable Diffusion v1.5 model with contrastive loss and regularization set as used in [2]. 'No attrib': No attribute editing; 'Dub chin': Double chin; 'Eyebrows': Bushy eyebrows; 'Mo open': Mouth slightly open.

We further compare local attribute editing with global attribute editing using t-SNE in Fig. 11. We observe that even our global attribute editing method is able to maintain evenly distributed identity embeddings across *bald*, *young* and *male* attributes thus, reflecting adherence to biometric features irrespective of different types of semantic and demographic attribute edits.

Comparison with GAN. We present additional comparisons with Talk-to-Edit [18], an interactive GAN-based facial editing technique. Talk-to-Edit provides fine-grained facial attribute editing via dialog interaction, similar to our approach. The method uses a language encoder to convert the user's request into an 'editing encoding' that encapsulates information about the degree and direction of change of the target attribute, and seeks user feedback to iteratively edit the desired attribute. The authors use the semantic field to preserve attribute localization and an identity-retention loss for maintaining identities. Talk-to-Edit works very well on synthetic images but struggles with real images. We observe that Talk-to-Edit can produce several artifacts (fails to add *Bangs* in case of black hair, degrades face when asked to increase the intensity of *Smiling*, fails to remove facial hair when asked with *No beard*, fails to detect presence of *Eyeglasses*, and defaults to the original for a majority of face images when asked to change age with *Young* attribute edit operation.
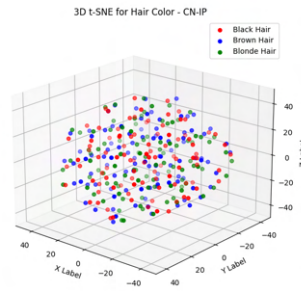
See Fig. 12 for examples of failure cases.
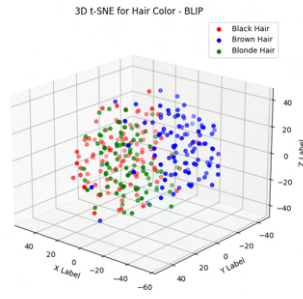
(a) BLIPDiffusion
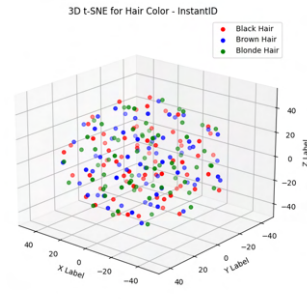
(b) InstantID

(c) CN-IP(ours)

Figure 10. Comparison of 3-D t-SNE plots for AdaFace embedding for BLIPDiffusion, InstantID and CN-IP (local). Identity embeddings are expected to be evenly distributed regardless of the hair color variations. However, we note that the ID embeddings from BLIPDiffusion images present relatively distinct pattern for different hair color explaining the loss in biometric fidelity compared to InstantID and CN-IP.
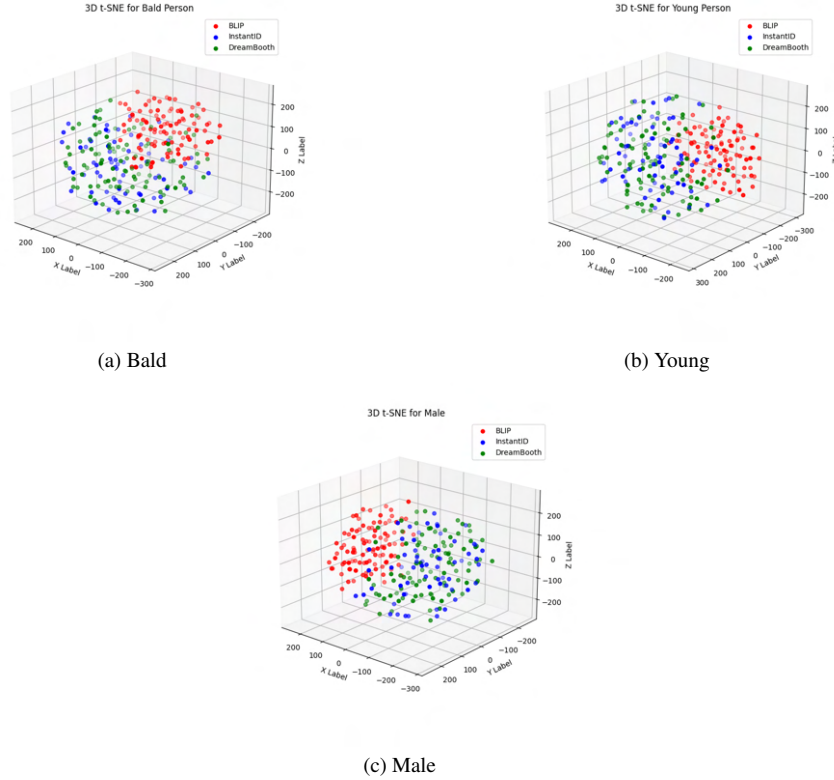
(a) Bald

(b) Young

(c) Male

Figure 11. Comparison of 3-D t-SNE plots for AdaFace embedding for BLIPDiffusion, InstantID and DB-proposed (global) on *bald, young* and *male* attributes. Identity embeddings are expected to be evenly distributed regardless of the variations in hair style, age and digital alteration of sexual cues. We note that our DB-prop. is able to retain such coarse details with better biometric fidelity compared to BLIPDiffusion. Note InstantID defaults to the original image in several cases resulting in identity preservation but fails to perform attribute editing.
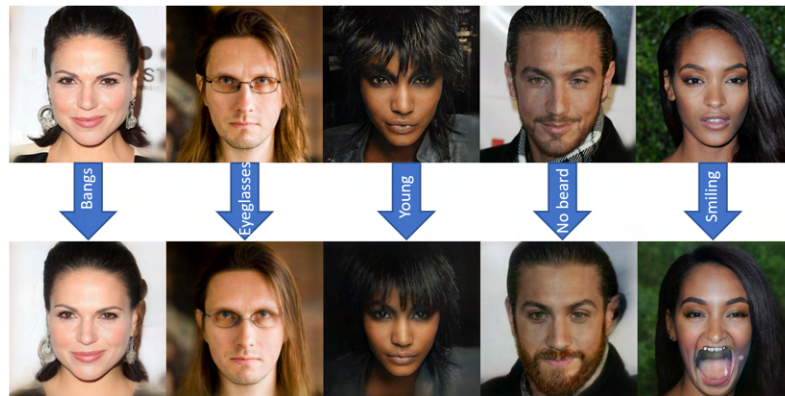

Figure 11. Failure cases of GAN-based facial attribute editing operation such as Talk-to-Edit [18]. In a majority of cases, it either fails to add attribute thus reconstructing the original image or produces unrealistic artifacts. In *Eyeglasses*, the method fails to detect glasses in the input image. So, it produces artifacts when trying to remove the glasses in the first edit direction. Similarly, in the *No beard* example, it adds facial hair.
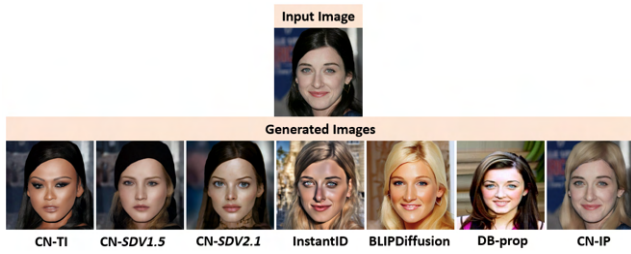
(a) Prompt: photo of a person with blond hair
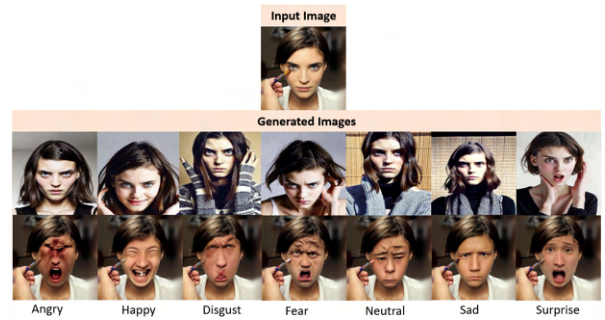
(b) Prompt(s): photo of a person with * expression.

Figure 12. Examples of failure cases particularly for expression editing using the local attribute editing method. We disabled the safety-checker option to present the failure cases. The CN-IP model does not produce outputs for expressions in a majority of cases.