
F?D: On understanding the role of deep feature spaces on face generation evaluation

Krish Kabra

Department of Electrical & Computer Engineering
Rice University
Houston, TX 77005
kk80@rice.edu

Guha Balakrishnan

Department of Electrical & Computer Engineering
Rice University
Houston, TX 77005
gb35@rice.edu

Abstract

Perceptual metrics, like the Fréchet Inception Distance (FID), are widely used to assess the similarity between synthetically generated and ground truth (real) images. The key idea behind these metrics is to compute errors in a deep feature space that captures perceptually and semantically rich image features. Despite their popularity, the effect that different deep features and their design choices have on a perceptual metric has not been well studied. In this work, we perform a causal analysis linking differences in semantic attributes and distortions between face image distributions to Fréchet distances (FD) using several popular deep feature spaces. A key component of our analysis is the creation of synthetic counterfactual faces using deep face generators. Our experiments show that the FD is heavily influenced by its feature space’s training dataset and objective function. For example, FD using features extracted from ImageNet-trained models heavily emphasize hats over regions like the eyes and mouth. Moreover, FD using features from a face gender classifier emphasize hair length more than distances in an identity (recognition) feature space. Finally, we evaluate several popular face generation models across feature spaces and find that StyleGAN2 consistently ranks higher than other face generators, except with respect to identity (recognition) features. This suggests the need for considering multiple feature spaces when evaluating generative models and using feature spaces that are tuned to nuances of the domain of interest.

1 Introduction

Rapid advances in generative image models such as variational autoencoders (VAEs) [1, 2], generative adversarial networks (GANs) [3–6], and diffusion models [7–10], point to a future where synthetic images play a significant role in society [11–13]. Therefore, it is crucial to continuously assess and improve how we evaluate the performances of these generative models [14]. In particular, synthesis evaluation metrics should capture several factors, including correlation to human perception, robustness to insignificant variations and noise, and sensitivity to domain-specific semantics.

The gold standard in evaluating image generation quality is human annotation [15], which can provide nuanced and interpretable perceptual feedback, but comes at the cost of money and time. The current standards in automated evaluation are deep perceptual metrics like perceptual loss (LPIPS) [16] and

Fréchet Inception Distance (FID) [17]. These metrics embed images in the activation space of a deep neural network trained on a general auxiliary task and compute distances in that feature space. Deep perceptual metrics correlate better with human evaluation than classical metrics (e.g., PSNR) [16, 18]. Unfortunately, the complexity of deep feature spaces also makes them opaque and hard to interpret. For example, deep feature spaces will emphasize certain attributes over others (as we demonstrate in our experiments), and can even be influenced by spurious features unrelated to the domain of interest [19]. Given that deep generative models are now typically competing with each other for less than 5 FID points, it is unclear what such differences mean semantically. When evaluating face generators, answering questions such as “What effect does an imbalanced generation of skin tones have on FID?” or, “What is the effect of consistent distortion of the eyes on FID?” is crucial in helping engineers better understand their evaluation metric, which ultimately will enable them to mitigate biases inherent in their models and/or improve generation quality.

In this work, we propose a strategy to causally evaluate the effect variations in domain-specific characteristics have on a perceptual evaluation metric using synthetic data. We focus on face generation, the most popular domain for image synthesis studies with many important societal implications in applications like face analysis/recognition [20], deepfakes [21], virtual avatars [22, 23], and even healthcare [24]. We consider two types of face manipulations: semantic attributes (e.g., hats, skin tone, hair length) and distortions (blur). We perform causal studies, using experimental interventions that manipulate a single characteristic of an image at a time, and measure their effects on an evaluation metric (in our experiments, Fréchet distance) across different image spaces. For semantic interventions, we use deep face generators to construct a dataset of synthetic face pairs that differ (approximately) by only a single feature of interest. For distortions, we apply blur to semantic facial regions inferred by a face segmentation model.

Using this synthetic data, we evaluated Fréchet distances (FD) in six deep feature spaces trained on both general-domain (ImageNet [25], CLIP [26]) and in-domain (face) datasets, and with fully-supervised, semi-supervised, and unsupervised objective functions. We find that the sensitivity of FD to semantic attributes depends heavily on the training dataset *and* objective function used to train the feature space. For example, feature spaces trained on ImageNet tend to overly emphasize accessories like hats and eyeglasses, while ignoring important facial semantics such as eyes, expression, geometry, and skin texture. Moreover, feature spaces trained on in-domain face datasets with contrastive learning objectives emphasize features of the skin (e.g. skin tone and texture) but ignore details related to hair and background. Finally, we evaluated popular deep generative face models using FD, precision, and recall [27] metrics computed over these feature spaces. Our findings show that while StyleGAN2 outperforms other generators on most feature spaces, in a face identity (recognition) feature space it has significantly worse recall and FD than a popular diffusion model (LDM), and worse precision than a 3D-aware GAN (EG3D).

The results of this study demonstrate that deep feature spaces have significant and unique biases over in-domain attributes due to both training data and objective functions. These biases should be understood by researchers during synthesis evaluation. In addition, our experiments with face generators demonstrate the importance of considering multiple feature spaces during evaluation, and particularly those tuned to crucial details for the domain of interest.

1.1 Related Work & Background

Deep generative model evaluation involves computing the similarity between generated and real image distributions. Metrics such as the Inception Score (IS) [28] and Fréchet Inception Distance (FID) [17] embed images into a lower-dimensional perceptual feature space derived from the final layers of a deep convolutional neural network (InceptionV3 [29]) trained on ImageNet [25].

FID is currently the *de facto* image generation evaluation metric. It assumes that two Inception-embedded image distributions are multivariate Gaussian, and computes the Fréchet distance (FD), otherwise known as 2-Wasserstein or Earth Mover’s distance [30]:

$$FD(\mu_1, \Sigma_1, \mu_2, \Sigma_2) = \|\mu_1 - \mu_2\|_2^2 + \text{Tr}\left(\Sigma_1 + \Sigma_2 - 2(\Sigma_1 \Sigma_2)^{\frac{1}{2}}\right), \quad (1)$$

where (μ_1, Σ_1) and (μ_2, Σ_2) are the sample means and covariances of the image set embeddings (i.e., real and generated images), and $\text{Tr}(\cdot)$ is the matrix trace.



Figure 1: **Sample images in our proposed counterfactual face dataset.** (a) The first column shows synthetically generated base faces with predefined “neutral” characteristics. We manipulate each face along different attribute of interest, shown in the remaining columns. (b) The first column (base) shows real faces from FFHQ [4]. Faces in the remaining columns are distorted versions of the base faces with blur added to specific semantic regions. Note that we show different faces for some of the eyes, eyeglasses, and hat examples because the corresponding base faces do not contain those regions (eyeglasses are considered to occlude and therefore remove the eye regions).

Several recent works have exposed limitations of FD and suggest alternate evaluation metrics [31]. One major drawback of FD is its high bias, requiring 50,000 samples or more to yield accurate estimates. To combat this, Bińkowski *et al.* [32] propose relaxing the Gaussian assumption using a polynomial kernel and computing the squared maximum mean discrepancy, whereas Chong *et al.* [33] propose estimating a bias-free FD by extrapolating the score using the fact that it is linear with respect to $\frac{1}{N}$. Another limitation of FD is its obfuscation of sample quality and variation. To circumvent this, Sajjadi *et al.* [34] and Kynkäänniemi *et al.* [27] propose separating these two components of evaluation into separate metrics: *precision*, which quantifies sample quality, and *recall*, which quantifies sample coverage.

More recently, FID has been shown to be limited by unwanted biases and sensitivities. Parmar *et al.* [35] show that FID is sensitive to imperceptible artifacts caused by image processing operations such as resizing, compression and quantization. Kynkäänniemi *et al.* [19] show that out-of-domain (ImageNet) features in the background of face images also strongly affect FID. Morozov *et al.* [36] advocate computing FD using self-supervised feature spaces, arguing that these features are more transferable and robust than those in Inception.

Our work is inspired by these studies and builds on them in several ways. First, our focus is a *causal* understanding of how *in-domain* facial semantics affect FD in several deep feature spaces. Second, we propose a novel strategy to perform this evaluation via synthesizing counterfactual image pairs with deep generative models. Third, we focus on one important domain (faces), which allows us to perform more nuanced semantic evaluations (e.g., how skin tone or nose distortions affects FD) than the general case considered by all previous studies. Finally, we also provide evaluations of several popular deep face generation models in these feature spaces, yielding insight into how these models compare to one another.

2 Methods

For a given deep feature space, our goal is to quantify the sensitivity of an evaluation metric to image characteristics. In our experiments, we focus on face images and FD, and so we describe our methods here in that context. We form two questions for a given feature space:

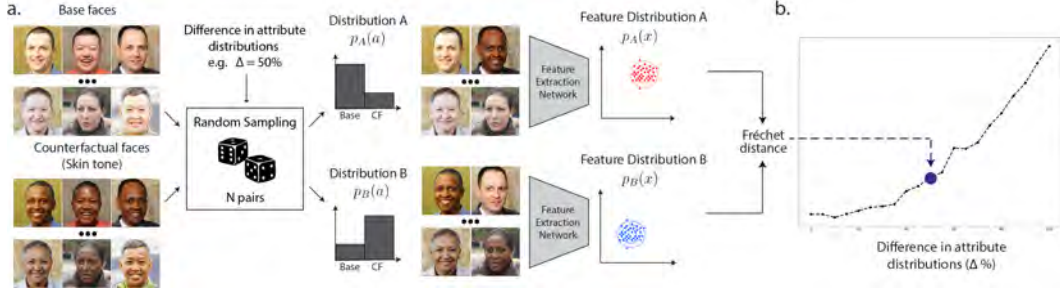


Figure 2: **Overview of our method for causally assessing the sensitivity of Fréchet distance to facial attributes.** a. We generate a set of base and counterfactual (CF) face pairs for an attribute or semantic distortion. In this figure, we use skin tone as an example. Given a difference in proportion for this attribute between two distributions ($\Delta \in [0\%, 100\%]$) and the number of faces per set (N), we construct two image sets A and B by randomly assigning base and CF faces to them such that this difference is achieved. We then extract the features for each image set using a pre-trained deep model (e.g., Inception, CLIP), and compute the FD between the two feature distributions. b. By creating set pairs for $\Delta \in [0\%, 100\%]$, we can generate a curve that summarizes the causal effect of a difference in attribute proportions on FD computed in a feature space.

1. How do differences between the semantic attribute distributions of two face image sets quantitatively affect FD?
2. How do distortions localized to a semantic region of a face quantitatively affect FD?

These questions align with two broad image characteristics that a generative model must capture: (1) semantic attributes for the domain, and (2) realistic details. Answering these questions requires causal reasoning, and ideally a *counterfactual* dataset consisting of pairs of faces that are identical except for a difference along one characteristic (i.e., semantic attribute or distortion). Real-face datasets contain significant attribute correlations [20] and are therefore not appropriate. Instead, we propose a synthetic approach. In the following sections, we outline the proposed methodology to construct synthetic data to answer each question¹. Example images synthesized by the proposed methods are shown in Figure 1.

2.1 Measuring the effect of semantic attribute differences on Fréchet distance

Consider two image sets A and B with feature distributions $p_A(x)$ and $p_B(x)$, where $x \in R^D$ is a feature space of an image. Furthermore, assume that A and B are identical except for a difference in their distributions over one semantic binary attribute with value $a \in \{0, 1\}$, which we denote by $p_A(a)$ and $p_B(a)$. Our goal is to quantify how the difference in attribute proportions between $p_A(a)$ and $p_B(a)$ (ranging from 0% when identical to 100% when completely dissimilar), affects $FD(\mu_A, \Sigma_A, \mu_B, \Sigma_B)$, the FD between $p_A(x)$ and $p_B(x)$. Figure 2 describes our analysis methodology to do so. We construct multiple sets of nearly identical faces using deep generative models (described below), each consisting of different proportions of values to a , and compute FD between the pairs to yield a curve summarizing causal effects (see Figure 2-right, and Figure 3).

This analysis requires the creation of sets A and B , counterfactual face sets that differ based on only a . We use a two-step process to create this data synthetically. First, we synthesize a set of *base faces* that exhibit predefined uniform characteristics of light skin tones and short hair, and no: facial hair, make-up, frowning expressions, hats, or eyeglasses. To do this, we obtained the face generation models of a previous facial causal benchmarking study [20] based on StyleGAN2 [5] trained on the Flickr-Faces-HQ (FFHQ) dataset [4] and orthogonalized linear latent space traversals (OLLT). We filter these faces via human evaluations to ensure they meet the defined criteria. In our experiments, we used a total of 1427 filtered base faces.

In the second step, we synthesize counterfactual pairs from the base faces for each attribute a . In our experiments, we analyzed 12 binary attributes corresponding to various facial semantics including geometry, skin tone, skin texture, hairstyle, and accessories. The attributes analyzed are shown by the

¹We will make our synthesized dataset publicly available.

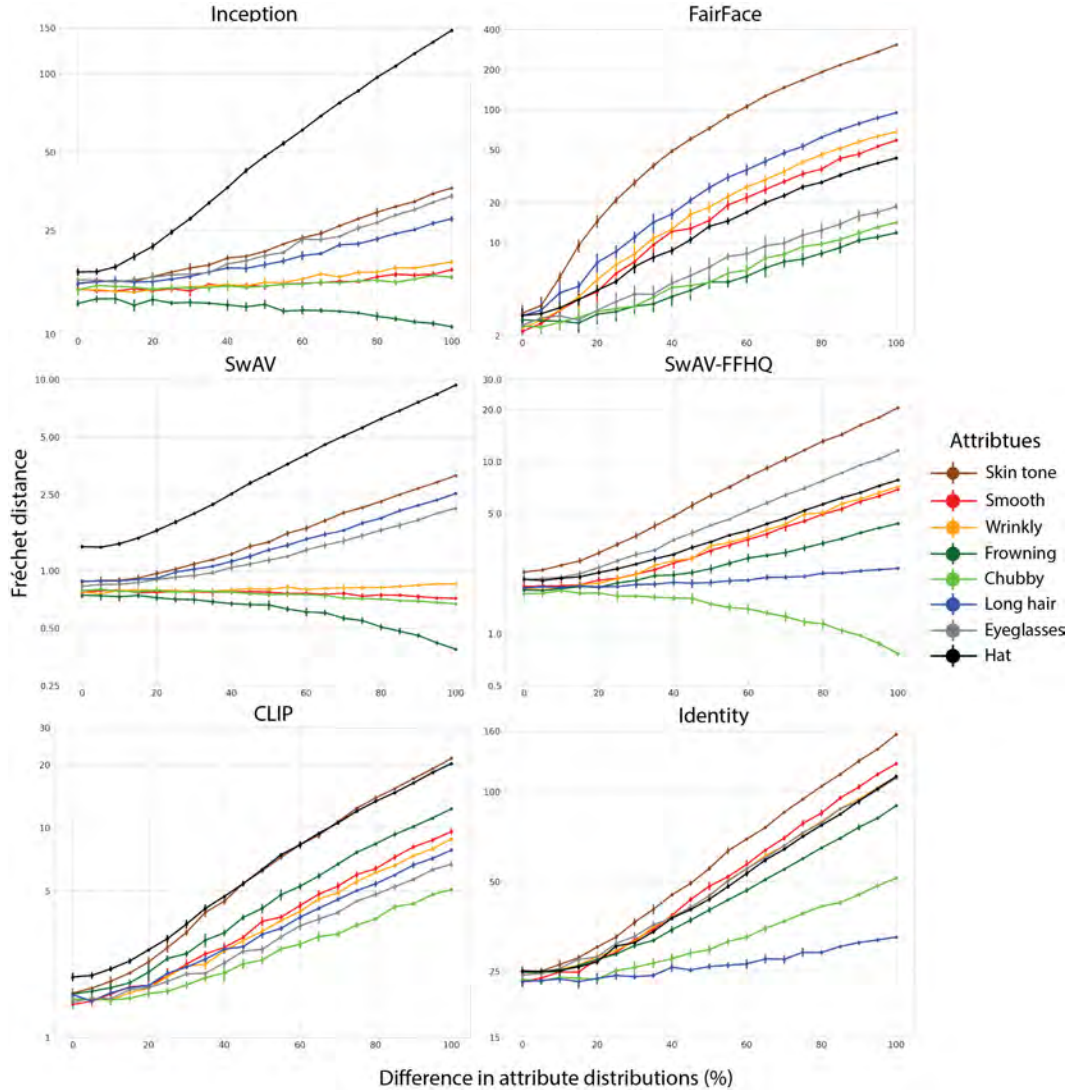


Figure 3: **Results for causal sensitivity analysis of Fréchet distances (FD) in different feature spaces with respect to semantic attributes.** For each percentage difference in attribute proportion (i.e. point along the x-axis), we sample 10 random draws of 1000 counterfactual face pairs to construct face sets, from which the FD in a feature space is computed, shown by the y-axis in log scale. See Figure 2 and Section 2.1 for a more detailed description of the approach. The three feature spaces on the left are trained on general-domain data, while the ones on the right are trained on in-domain (face) data. Each feature space under- or over-emphasizes certain attributes based on its training dataset and objective functions. Note that absolute values of FD are not meaningful to compare across feature spaces due to arbitrary scaling differences.

columns in Figure 1a. We utilize one of three different image manipulation methods based on the attribute type: (1) OLLT, (2) StyleCLIP [37], and (3) image inpainting with Stable Diffusion [38]. We choose the best method for each attribute based on a qualitative assessment of how well each method can manipulate the attribute while holding others constant. We show some example attribute counterfactuals in Figure 1a. We provide a complete account of models, experimental parameters, and details used to create the synthetic dataset in Supplementary.

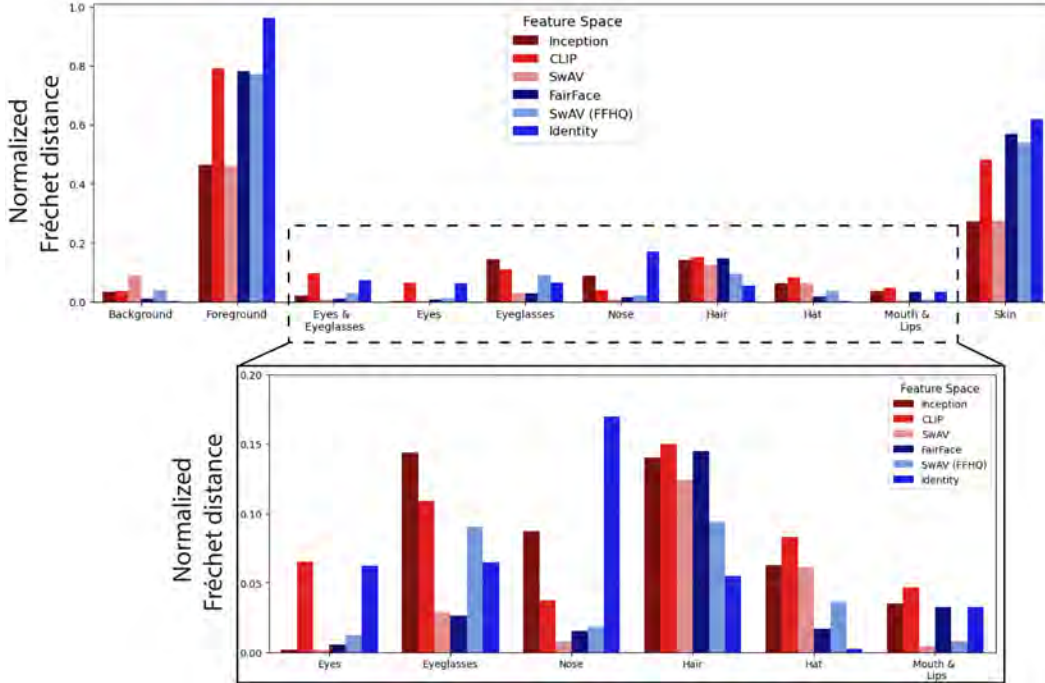


Figure 4: **Results for semantic region distortion (blur) analysis.** (Top) Bar plot comparing the normalized Fréchet distances per semantic region blur. We normalize distances in each feature space by dividing by the distance between the original and fully blurred image set (“All” in Figure 3) in that space. (Bottom) Zoomed-in plot to clearly visualize results for semantic regions that occupy less than 25% of the face image on average.

2.2 Measuring the effect of blurring semantic regions on Fréchet distance

The purpose of this analysis is to understand how a systematic distortion outputted by a face generator for a specific semantic region (e.g., nose, hair) impacts FD. In our experiments, we focused on heavy blur (see Figure 1, though many others may also be explored [16]). For each region, we use real FFHQ face images that contain that region (accessories like hats and eyeglasses are not in every image) as one distribution (set A), and apply Gaussian blur to these images *only in that region* using segmentation masks obtained from a public face segmentation model² (set B). We considered 9 regions in our experiments, as shown in Figure 1b. For our analysis, we simply report FD with respect to distorting each semantic region (see Figure 4).

3 Experiments

We conduct our analyses using six deep feature spaces with publicly available parameters (a list of model and parameter URLs are in Supplementary):

1. **Inception:** V3 model trained on the ILSVRC-2012 (ImageNet) dataset for classification [29].
2. **CLIP:** ViT-B/32 model trained on OpenAI’s proprietary image-text pair dataset using a contrastive loss [26].
3. **SwAV:** ResNet-50 model trained on the ILSVRC-2012 (ImageNet) for unsupervised learning of visual features by contrasting cluster assignments [39].
4. **FairFace:** ResNet-34 model trained on FairFace for race, age, and gender classification [40].
5. **SwAV-FFHQ:** ResNet-50 model trained on FFHQ for unsupervised learning of facial features by contrasting cluster assignments [39, 41].

²<https://github.com/zllrunning/face-parsing.PyTorch>

6. **Identity:** ResNet-34 model trained on Glint360k for facial recognition with a contrastive loss [42].

The first three feature spaces are trained on general-domain (non-face) data, while the last three are trained on faces only.

3.1 Causal sensitivity of feature spaces to image characteristics

We first present sensitivity analyses of FD with respect to facial attribute proportions. Figure 3 presents our results. Our total dataset used for this analysis contains 1427 counterfactual face pairs. For each percentage difference in attribute proportion (i.e. point along the x-axis), we sample 10 random draws of 1000 pairs to construct face sets, from which the FD in a feature space is computed. For clarity, we plotted only 8 attributes (full results for all 12 attributes are in Supplementary). The points and error bars shown in the plots correspond to the mean and standard deviation respectively. A direct comparison of FD values across feature spaces is not meaningful, as the scale of distances vary across features. However, the difference in trends between the attribute curves may be compared across plots. For example, Inception and SwAV clearly emphasize hats with respect to other feature spaces, while FairFace and SwAV-FFHQ emphasize skin tone.

Next, we present sensitivity analyses of FD with respect to localized distortions. In order to compare FD across different feature spaces, we normalize distances by dividing them by the distance between the original and entirely blurry images (“all” category in Figure 1) in that feature space. We show unnormalized FD for each feature space in Supplementary. There are significant differences in how each feature space is affected by regional distortions. For example, Inception and SwAV are virtually unaffected by the eyes compared to other spaces, and Identity is most affected by the nose.

3.2 Analysis of face generators in different feature spaces

Next, we present evaluations of four popular, publicly available face generation models using metrics computed in each feature space: StyleGAN2 [5], EG3D [6], latent diffusion model (LDM) [9], and Nouveau variational autoencoder (NVAE) [2]. For StyleGAN2 and EG3D, we evaluate the models both with and without truncation [43, 44] ($\psi = 0.7$, truncation cutoff = 14). We evaluate models using FD and k -nearest neighbors precision and recall metrics [27]. These precision and recall measures approximate sample quality (realism) and sample coverage, respectively. We use the entire FFHQ dataset (70,000 images) and 50,000 samples from each generative model. Complete results are shown in Table 1.

4 Discussion

In this section, we discuss key observations from our experiments, followed by limitations.

Feature spaces learned using ImageNet under-emphasize important facial semantics regardless of the training objective. Figure 3 illustrates that FD in feature spaces learned using ImageNet (Inception and SwAV) are most sensitive to differences in the proportion of hats, consistent with findings from Kynkäänniemi *et al.* [19]. However, interestingly, the FD computed using SwAV features are also sensitive to hats, even though that model is not explicitly trained to classify ImageNet classes. This is reasonable since self-supervised learning is known to be an effective pretraining strategy for ImageNet classification.

The plots also illustrate that FD computed using ImageNet-learned spaces are highly insensitive to distributional differences in skin texture (“wrinkly” and “smooth”), geometry (“chubby”), and expression (“frowning”). This result is also affected by a subtle interplay between the mean and trace terms of the FD in Equation (1)(see Supplementary). As the two distributions become more skewed in our sensitivity analyses (towards 0 or 100 % in Figure 3), the distribution means become more dissimilar, but their variances also decrease and reduce the trace term. This suggests another challenge in using FD alone: they can obfuscate differences in distribution modes versus distribution shapes.

Figure 4 shows that FD computed using Inception and SwAV spaces are insensitive to the blurring of the eyes, and SwAV is insensitive to the blurring of the nose and mouth. This shows that systematic

Table 1: **Generative model evaluation using different deep image spaces and metrics.** We evaluate 50K images synthesized by each generative models with respect to the full FFHQ (70K) dataset. For each feature space, we highlight the top three performing models with the following key: **First**, **Second**, **Third**. Note that FD values are not meaningful to compare across feature spaces due to arbitrary scaling differences. StyleGAN2 generally outperforms all other models, but in Identity space is worse in Fréchet distance and Recall than LDM, and worse in Precision than EG3D.

Fréchet Distance (\downarrow)						
	Inception	CLIP	SwAV	FairFace	SwAV (FFHQ)	Identity
StyleGAN2 [5]	3.1	1.8	0.6	1.6	0.4	17.8
StyleGAN2 (Truncated)	21.0	8.2	2.0	27.1	4.0	61.2
EG3D [6]	16.5	7.0	2.1	34.2	9.2	162.0
EG3D (Truncated)	40.2	13.0	3.3	41.8	16.7	221.3
LDM [9]	10.0	3.6	1.7	10.9	1.4	6.9
NVAE [2]	35.9	9.7	5.4	56.9	5.8	44.1

Precision (%) (\uparrow)						
	Inception	CLIP	SwAV	FairFace	SwAV (FFHQ)	Identity
StyleGAN2	67.4	77.0	79.1	84.5	74.3	59.4
StyleGAN2 (Truncated)	83.3	89.0	89.8	88.7	67.7	88.0
EG3D	67.1	61.7	55.3	63.2	48.5	86.0
EG3D (Truncated)	79.8	82.8	72.1	71.1	38.6	92.8
LDM	72.2	72.0	74.7	85.6	78.8	37.8
NVAE	65.3	57.7	69.5	82.3	49.0	65.5

Recall (%) (\uparrow)						
	Inception	CLIP	SwAV	FairFace	SwAV (FFHQ)	Identity
StyleGAN2	50.2	42.3	25.1	81.9	79.5	3.5
StyleGAN2 (Truncated)	26.5	14.8	7.1	61.1	66.3	0.7
EG3D	26.9	20.3	9.7	80.7	20.9	0.0
EG3D (Truncated)	11.1	5.6	1.7	49.8	11.9	0.0
LDM	38.5	38.5	10.6	77.8	71.3	22.8
NVAE	12.1	10.4	0.4	55.7	46.5	1.9

degradations to the eyes, nose, or mouth, will not impact the FD in ImageNet-based feature spaces. Generative model designers should pay extra attention to these semantic “blind spots.”

The training objective influences which facial semantics are emphasized by a deep feature space. Figure 3 shows that while in-domain feature spaces (FairFace, SwAV-FFHQ, Identity) are all highly sensitive to differences in skin tone, skin texture, and facial accessories, there do exist several notable dissimilarities. For example, FairFace is far more sensitive to hair length, compared to SwAV-FFHQ and Identity. This is further supported by the relatively small effect that blurring the hair has on SwAV-FFHQ and Identity compared to FairFace. Another notable distinction is that both FairFace and SwAV-FFHQ fail to capture distortions localized to the eyes, nose, mouth, and lips, whereas Identity does. We speculate that these differences are a consequence of the feature spaces capturing semantic characteristics that pertain most to the objective function used during training. FairFace is trained to classify perceived gender, which is correlated with hair length. On the other hand, Identity is trained to match faces corresponding to the same person, which should be invariant to hairstyle and hair length. SwAV is trained to match cropped views of an image, for which hair length is likely not a robust feature. Therefore, we suggest that generative model designers should not naïvely expect in-domain feature spaces to be sensitive to all domain-specific semantics. Rather, we advocate carefully considering how the training objective may influence features, and empirically investigating these sensitivities.

Image-language models trained on massive general datasets capture many important semantic characteristics of faces. The sensitivity analyses for both semantic attributes and distortions show the CLIP features are sensitive to all studied characteristics. In particular, CLIP provides a significant FD for all distorted facial region irrespective of the size of the region in pixels. This is likely because of two reasons: (1) CLIP is trained on a massive dataset, and (2) text provides a rich source of information on perceptual features to the image encoder that cannot have otherwise been learned using classical supervision. Based on these results, we encourage generative model designers to move away from perceptual features extracted from models trained on ImageNet (Inception, VGG [45], SwAV) and use large image-language models like CLIP.

StyleGAN2 consistently outperforms other face generators, except with respect to identities. Table 1 shows that across almost all feature spaces and evaluation metrics, StyleGAN2 outperforms other face generators. However, when using Identity space, LDM and EG3D outperform StyleGAN2 in different ways. We speculate that because EG3D is trained to be aware of 3D geometry, it models features like the nose and eyes (which are important in Identity space as indicated by Figure 4 well, yielding higher Precision. LDM may be able to capture a wider diversity of identities compared to GAN models because it is not susceptible to the mode collapse issues known to plague GANs [3], yielding higher Recall and FD. This observation is further reinforced by the identity metrics for NVAE in comparison to EG3D. Although NVAE generates samples less precise in identity than StyleGAN2 and EG3D, it dethrones the GAN models for their third-place spots for Recall and FD.

4.1 Limitations

Our causal analysis of semantic attributes assumes perfectly counterfactual face pairs. However, it is not possible to exactly isolate one attribute from others when working with deep generators due to the correlations the generator learns from its training distribution. For example, when manipulating eyeglasses, we find that the skin texture becomes more wrinkly due to correlations between age and the propensity to wear eyeglasses. Moreover, in the case of synthesizing wrinkly faces, we find that the manipulation also causes the face to have squinted eyes. In general however, such correlations are known to be even more dramatic in real datasets [20], which makes synthetic generation a more attractive option for such analysis.

Our semantic attribute analysis uses a sample size of 1000 images per set, which results in biased FD estimates [32, 33]. However, given that sample size was consistent throughout the experiment, the trend and shapes of curves shown in Figure 3 are accurate. This sample size bias is also a factor in our analysis for distortions of semantic regions such as hats and eyeglasses which are present in only a smaller subset (~ 10,000) of the entire FFHQ dataset.

We used publicly available deep networks for our feature spaces, but these networks vary in architecture type, size, and final layer feature number (for example, Inception and SwAV features have 2048 dimensions, while others have 512). These factors could be potential confounders in our causal analysis, though we believe that training set composition and objective functions likely have a far stronger influence on our results.

5 Conclusion

In this work, we proposed a strategy to causally evaluate the impact of variations in domain-specific characteristics on generative evaluation metrics using synthetic data in the context of face generation. We present a thorough study of the sensitivities FD computed using several deep feature spaces has with respect to facial attributes. Moreover, we provide an analysis of popular deep generative models evaluated in those feature spaces. The results of this study demonstrate that deep feature spaces have significant and unique biases over in-domain attributes due to both training data and objective functions.

In summary, biases of a given feature space should be fully understood by researchers before using them for synthesis evaluation. Our experiments with face generators demonstrate the importance of considering multiple feature spaces during evaluation – particularly those tuned to crucial details for the domain of interest (like identity for faces) – to get a full picture of a model’s strengths and shortcomings. As image generation models continue to improve at a rapid pace, such careful

evaluation is necessary to make meaningful progress, including mitigating biases and enhancing the overall quality of generative image models.

6 Broader impacts

Improving the understanding of image generation evaluation metrics will enable researchers and developers to assess the performance of different generative models more effectively. This, in turn, promotes the development of more accurate, realistic, and reliable generation systems. Evaluation metrics also play a crucial role in mitigating biases of face generation models. Nevertheless, we recognize that face-generation technologies have far-reaching implications, including potentially harmful applications involving deepfakes and identity manipulation. Therefore, improving evaluation procedures of such methods can consequently increase the risks and harms associated with AI-generated facial content.

References

- [1] Ali Razavi, Aaron van den Oord, and Oriol Vinyals. Generating Diverse High-Fidelity Images with VQ-VAE-2. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.
- [2] Arash Vahdat and Jan Kautz. NVAE: A Deep Hierarchical Variational Autoencoder. In *Advances in Neural Information Processing Systems*, volume 33, pages 19667–19679. Curran Associates, Inc., 2020.
- [3] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Commun. ACM*, 63(11):139–144, October 2020.
- [4] Tero Karras, Samuli Laine, and Timo Aila. A Style-Based Generator Architecture for Generative Adversarial Networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4401–4410, 2019.
- [5] Tero Karras, Samuli Laine, Miika Aittala, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. Analyzing and Improving the Image Quality of StyleGAN. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 8110–8119, 2020.
- [6] Eric R. Chan, Connor Z. Lin, Matthew A. Chan, Koki Nagano, Boxiao Pan, Shalini De Mello, Orazio Gallo, Leonidas J. Guibas, Jonathan Tremblay, Sameh Khamis, Tero Karras, and Gordon Wetzstein. Efficient Geometry-Aware 3D Generative Adversarial Networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16123–16133, 2022.
- [7] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising Diffusion Probabilistic Models. In *Advances in Neural Information Processing Systems*, volume 33, pages 6840–6851. Curran Associates, Inc., 2020.
- [8] Yang Song, Conor Durkan, Iain Murray, and Stefano Ermon. Maximum Likelihood Training of Score-Based Diffusion Models. In *Advances in Neural Information Processing Systems*, November 2021.
- [9] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-Resolution Image Synthesis With Latent Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022.
- [10] Prafulla Dhariwal and Alexander Nichol. Diffusion Models Beat GANs on Image Synthesis. In *Advances in Neural Information Processing Systems*, volume 34, pages 8780–8794. Curran Associates, Inc., 2021.
- [11] Jie Gui, Zhenan Sun, Yonggang Wen, Dacheng Tao, and Jieping Ye. A review on generative adversarial networks: Algorithms, theory, and applications. *IEEE Transactions on Knowledge and Data Engineering*, 35(4):3313–3332, 2023.

- [12] Tonghe Wang, Yang Lei, Yabo Fu, Jacob F. Wynne, Walter J. Curran, Tian Liu, and Xiaofeng Yang. A review on medical imaging synthesis using deep learning and its clinical applications. *Journal of Applied Clinical Medical Physics*, 22(1):11–36, 2021.
- [13] Daniel McDuff, Theodore Curran, and Achuta Kadambi. Synthetic data in healthcare, 2023.
- [14] Ryan Burnell, Wout Schellaert, John Burden, Tomer D. Ullman, Fernando Martinez-Plumed, Joshua B. Tenenbaum, Danaja Rutar, Lucy G. Cheke, Jascha Sohl-Dickstein, Melanie Mitchell, Douwe Kiela, Murray Shanahan, Ellen M. Voorhees, Anthony G. Cohn, Joel Z. Leibo, and Jose Hernandez-Orallo. Rethink reporting of evaluation results in AI. *Science*, 380(6641):136–138, April 2023.
- [15] Sharon Zhou, Mitchell Gordon, Ranjay Krishna, Austin Narcomey, Li F Fei-Fei, and Michael Bernstein. HYPE: A Benchmark for Human eYe Perceptual Evaluation of Generative Models. In *Advances in Neural Information Processing Systems*, volume 32, 2019.
- [16] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The Unreasonable Effectiveness of Deep Features as a Perceptual Metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 586–595, 2018.
- [17] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.
- [18] Qiantong Xu, Gao Huang, Yang Yuan, Chuan Guo, Yu Sun, Felix Wu, and Kilian Weinberger. An empirical study on evaluation metrics of generative adversarial networks. *arXiv preprint arXiv:1806.07755*, 2018.
- [19] Tuomas Kynkäänniemi, Tero Karras, Miika Aittala, Timo Aila, and Jaakko Lehtinen. The Role of ImageNet Classes in Fréchet Inception Distance. In *The Eleventh International Conference on Learning Representations*, February 2023.
- [20] Guha Balakrishnan, Yuanjun Xiong, Wei Xia, and Pietro Perona. Towards Causal Benchmarking of Bias in Face Analysis Algorithms. In Nalini K. Ratha, Vishal M. Patel, and Rama Chellappa, editors, *Deep Learning-Based Face Analytics*, Advances in Computer Vision and Pattern Recognition, pages 327–359. Springer International Publishing, Cham, 2021.
- [21] Ruben Tolosana, Ruben Vera-Rodriguez, Julian Fierrez, Aythami Morales, and Javier Ortega-Garcia. Deepfakes and beyond: A Survey of face manipulation and fake detection. *Information Fusion*, 64:131–148, 2020.
- [22] Justin N. M. Pinkney and Doron Adler. Resolution dependent gan interpolation for controllable image synthesis between domains, 2020.
- [23] Rameen Abdal, Hsin-Ying Lee, Peihao Zhu, Menglei Chai, Aliaksandr Siarohin, Peter Wonka, and Sergey Tulyakov. 3davargan: Bridging domains for personalized editable avatars, 2023.
- [24] Zhen Wang, Yunhao Ba, Pradyumna Chari, Oyku Deniz Bozkurt, Gianna Brown, Parth Patwa, Niranjan Vaddi, Laleh Jalilian, and Achuta Kadambi. Synthetic Generation of Face Videos With Plethysmograph Physiology. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 20587–20596, 2022.
- [25] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. ImageNet: A large-scale hierarchical image database. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 248–255, June 2009.
- [26] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763. PMLR, July 2021.

- [27] Tuomas Kynkäänniemi, Tero Karras, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Improved Precision and Recall Metric for Assessing Generative Models. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.
- [28] Tim Salimans, Ian Goodfellow, Wojciech Zaremba, Vicki Cheung, Alec Radford, Xi Chen, and Xi Chen. Improved Techniques for Training GANs. In *Advances in Neural Information Processing Systems*, volume 29. Curran Associates, Inc., 2016.
- [29] Christian Szegedy, Vincent Vanhoucke, Sergey Ioffe, Jon Shlens, and Zbigniew Wojna. Rethinking the Inception Architecture for Computer Vision. In *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 2818–2826, June 2016.
- [30] D. C Dowson and B. V Landau. The Fréchet distance between multivariate normal distributions. *Journal of Multivariate Analysis*, 12(3):450–455, September 1982.
- [31] Ali Borji. Pros and cons of GAN evaluation measures: New developments. *Computer Vision and Image Understanding*, 215:103329, January 2022.
- [32] Mikołaj Bińkowski, Danica J. Sutherland, Michael Arbel, and Arthur Gretton. Demystifying MMD GANs. In *International Conference on Learning Representations*, May 2023.
- [33] Min Jin Chong and David Forsyth. Effectively Unbiased FID and Inception Score and Where to Find Them. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6070–6079, 2020.
- [34] Mehdi S. M. Sajjadi, Olivier Bachem, Mario Lucic, Olivier Bousquet, and Sylvain Gelly. Assessing generative models via precision and recall. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.
- [35] Gaurav Parmar, Richard Zhang, and Jun-Yan Zhu. On Aliased Resizing and Surprising Subtleties in GAN Evaluation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11410–11420, 2022.
- [36] Stanislav Morozov, Andrey Voynov, and Artem Babenko. On Self-Supervised Image Representations for GAN Evaluation. In *International Conference on Learning Representations*, January 2021.
- [37] Or Patashnik, Zongze Wu, Eli Shechtman, Daniel Cohen-Or, and Dani Lischinski. StyleCLIP: Text-Driven Manipulation of StyleGAN Imagery. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2085–2094, 2021.
- [38] Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davaadorj, and Thomas Wolf. Diffusers: State-of-the-art diffusion models, May 2023.
- [39] Mathilde Caron, Ishan Misra, Julien Mairal, Priya Goyal, Piotr Bojanowski, and Armand Joulin. Unsupervised Learning of Visual Features by Contrasting Cluster Assignments. In *Advances in Neural Information Processing Systems*, volume 33, pages 9912–9924. Curran Associates, Inc., 2020.
- [40] Kimmo Kärkkäinen and Jungseock Joo. FairFace: Face Attribute Dataset for Balanced Race, Gender, and Age for Bias Measurement and Mitigation. In *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1547–1557, January 2021.
- [41] Dmitry Baranchuk, Andrey Voynov, Ivan Rubachev, Valentin Khrukov, and Artem Babenko. Label-Efficient Semantic Segmentation with Diffusion Models. In *International Conference on Learning Representations*, January 2022.
- [42] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou. ArcFace: Additive Angular Margin Loss for Deep Face Recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 4690–4699, 2019.
- [43] Marco Marchesi. Megapixel size image creation using generative adversarial networks, 2017.

- [44] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. In *International Conference on Learning Representations*, 2019.
- [45] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015.
- [46] Camillo Lugaresi, Jiuqiang Tang, Hadon Nash, Chris McClanahan, Esha Uboweja, Michael Hays, Fan Zhang, Chuo-Ling Chang, Ming Guang Yong, Juhyun Lee, Wan-Teh Chang, Wei Hua, Manfred Georg, and Matthias Grundmann. MediaPipe: A Framework for Building Perception Pipelines, June 2019.

A Implementation details

A.1 Open-source models

We make use of many publically available, open-source codes, models, and parameters (checkpoints) for our work. Table A.1 summarizes the models, code repositories, and checkpoint links used in our implementation.

Table A.1: Summary of open-source codes, models, and parameters used in the implementation.

Model	Type	Code repository	Model checkpoint
StyleGAN2	FFHQ (1024×1024)	https://github.com/NVlabs/stylegan2-ada-pytorch	https://nvlabs-fi-cdn.nvidia.com/stylegan2-ada-pytorch/pretrained/ffhq.pkl
StyleCLIP		https://github.com/orpatashnik/StyleCLIP	
Stable Diffusion	v2 (Inpainting)	https://github.com/huggingface/diffusers	https://huggingface.co/stabilityai/stable-diffusion-2-inpainting
Face segmentor	BiSeNet (CelebAMask-HQ)	https://github.com/zllrunning/face-parsing.PyTorch	https://drive.google.com/open?id=154JgKpzCPW82qINcVieuPH3fZ2e0P812
InceptionV3		https://github.com/NVlabs/stylegan2-ada-pytorch	https://nvlabs-fi-cdn.nvidia.com/stylegan2-ada-pytorch/pretrained/metrics/inception-2015-12-05.pt
SwAV	ResNet-50 (800 epochs, batch size 4096)	https://github.com/facebookresearch/swav	https://dl.fbaipublicfiles.com/deepcluster/swav_800ep_pretrain.pth.tar
CLIP	ViT-B/32	https://github.com/openai/CLIP	https://openaipublic.azureedge.net/clip/models/40d365715913c9da98579312b702a82c18be219cc2a73407c4526f58eba950af/ViT-B-32.pt
FairFace	ResNet-34 (7 race)	https://github.com/dchen236/FairFace	https://drive.google.com/file/d/113QMzQzkBDmYMs9LwzvD-jxEZdBQ5J4X
SwAV-FFHQ	ResNet-50 (400 epochs, batch size 2048)	https://github.com/facebookresearch/swav	https://storage.yandexcloud.net/yandex-research/ddpm-segmentation/models/swav_checkpoints/ffhq.pth
Identity	ResNet-34 (Glint360k)	https://github.com/deepinsight/insightface	https://1drv.ms/u/s!AswpsD02toNKq01WY69vN58GR6mw?e=p90v5d

A.2 Counterfactual dataset: attributes

We use a two-step process to create our counterfactual facial attribute dataset. We first synthesize a set of *base faces* that exhibit predefined uniform characteristics of light skin tones and short hair, and no: facial hair, make-up, frowning expressions, hats, or eyeglasses. To accomplish this, we sample a set of intermediate-style latent vectors $\{\mathbf{w}_i : \mathbf{w}_i \in \mathcal{W}\}$. We then use orthogonalized linear latent space traversals (OLLT) to traverse the latent vectors in a direction corresponding to light skin tone and short hair³. Finally, we filter these faces via human evaluations to ensure they meet the defined criteria. The final number of base faces contained in the dataset amounted to 1427 images.

In the second step, we synthesize counterfactual pairs from the base faces for each of the 12 binary attributes analyzed in our experiments (see first column of Table A.2). To accomplish this, we utilize one of three different image manipulation methods based on the attribute type: (1) OLLT, (2) StyleCLIP [37], and (3) image inpainting with Stable Diffusion [38]. We choose the best method for each attribute based on a qualitative assessment of how well each method can manipulate the attribute while holding others constant. A summary of the manipulation method and parameters used for each

³The hyperplane coefficients for age, gender, hair length, and skin tone attributes were graciously provided by the authors upon request.

Table A.2: Implementation parameters for our counterfactual attribute synthesis approach per semantic face attribute.

Attribute	Method	Text prompt	Manipulation parameters
Hat	Stable Diffusion inpainting	"A photo of a face with a hat"	Guidance scale = 0.75
Eyeglasses	StyleCLIP	Neutral: "face"	$\alpha = 10$
		Target: "face with eyeglasses"	$\beta = 0.13$
Skin tone	OLLT	N/A	Step size = 0.5
Make-up	StyleCLIP	Neutral: "face"	$\alpha = 3$
		Target: "face with makeup"	$\beta = 0.12$
Wrinkly	StyleCLIP	Neutral: "face with skin"	$\alpha = 3$
		Target: "face with wrinkly skin"	$\beta = 0.09$
Smooth	StyleCLIP	Neutral: "face with skin"	$\alpha = -3$
		Target: "face with wrinkly skin"	$\beta = 0.09$
Chubby	StyleCLIP	Neutral: "face"	$\alpha = 5$
		Target: "chubby face"	$\beta = 0.25$
Slim	StyleCLIP	Neutral: "face"	$\alpha = -5$
		Target: "chubby face"	$\beta = 0.25$
Slim	StyleCLIP	Neutral: "face"	$\alpha = -5$
		Target: "chubby face"	$\beta = 0.25$
Frowning	StyleCLIP	Neutral: "smiling face"	$\alpha = 5$
		Target: "frowning face"	$\beta = 0.20$
Hair length	StyleCLIP	Neutral: "face with hair"	$\alpha = 15$
		Target: "face with long hair"	$\beta = 0.20$
Curly	StyleCLIP	Neutral: "face with hair"	$\alpha = 5$
		Target: "face with curly hair"	$\beta = 0.25$
Fringe	StyleCLIP	Neutral: "face with hair"	$\alpha = 5$
		Target: "face with fringe hair"	$\beta = 0.15$

attribute is listed in Table A.2. To manipulate skin tone, we use OLLT to traverse in the direction of dark skin tones. For wearing a hat, we first automatically mask out a region reaching from the bottom of the forehead to the top of the image using 3D facial landmarks detected by MediaPipe face mesh model [46]. We then performed image inpainting using Stable Diffusion with the prompt "a photo of a face with a hat". For all other attributes, we use StyleCLIP to traverse along a direction that corresponds to the text prompts detailed in Table A.2. Note that for some attributes, namely "slim" and "smooth", we traverse in the negative direction of the text prompt. We experimentally found that these attributes are best manipulated by traversing in these negative directions as opposed to the corresponding positive directions (e.g. "slim face" and "face with smooth skin").

A.3 Counterfactual dataset: distortions (blur)

To create our counterfactual distortions (blur) dataset, we apply heavy blur to 9 semantic regions on real FFHQ face images. The regions for each image are obtained using segmentation masks obtained from a public face segmentation model (see Table A.1). The heavy blur is defined as a Gaussian blur with kernel size of 111×111 pixels and standard deviation $\sigma = 100$ pixels applied to a 512×512 image. The counterfactuals are synthesized by replacing the region of interest in the real image with the corresponding region in the blurred image.

B Additional experimental results

In this section, we present the full results for the causal sensitivity analyses of Fréchet distance (FD) in all 6 feature spaces to image characteristics. Figures B.1 to B.6 plot the FD against differences in facial attribute proportions across all 12 attributes analyzed. Additionally, the mean and trace terms that contribute to the total FD are shown. Figure B plots the (unnormalized) FD for each feature space when the specified semantic region is heavily blurred.

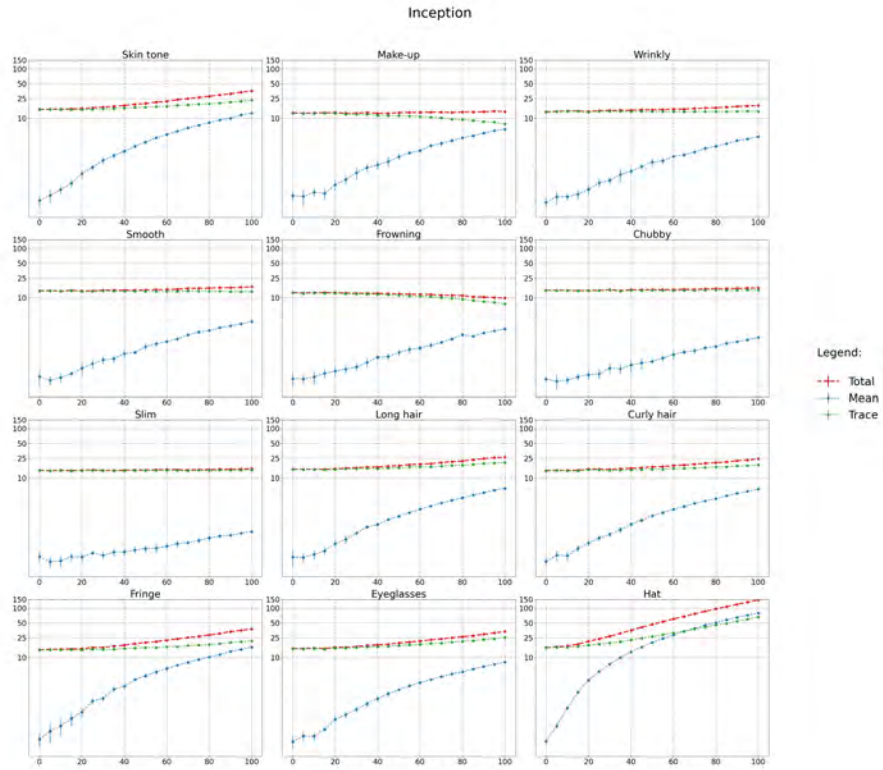


Figure B.1: Results for causal sensitivity analysis of Fréchet distances in the Inception feature space.

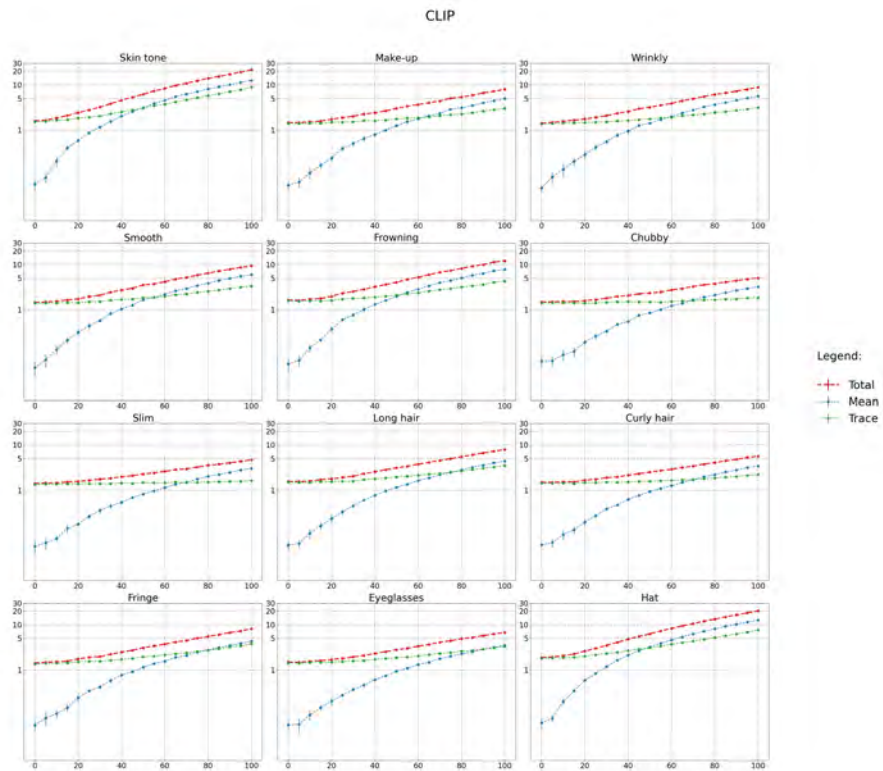


Figure B.2: Results for causal sensitivity analysis of Fréchet distances in the CLIP feature space.

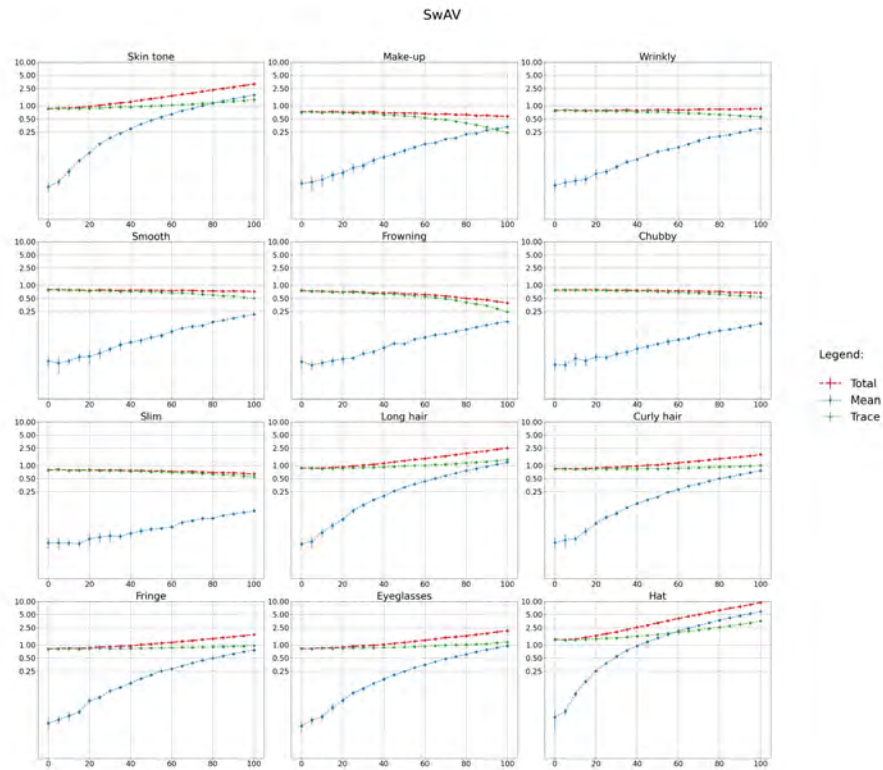


Figure B.3: Results for causal sensitivity analysis of Fréchet distances in the SwAV feature space.

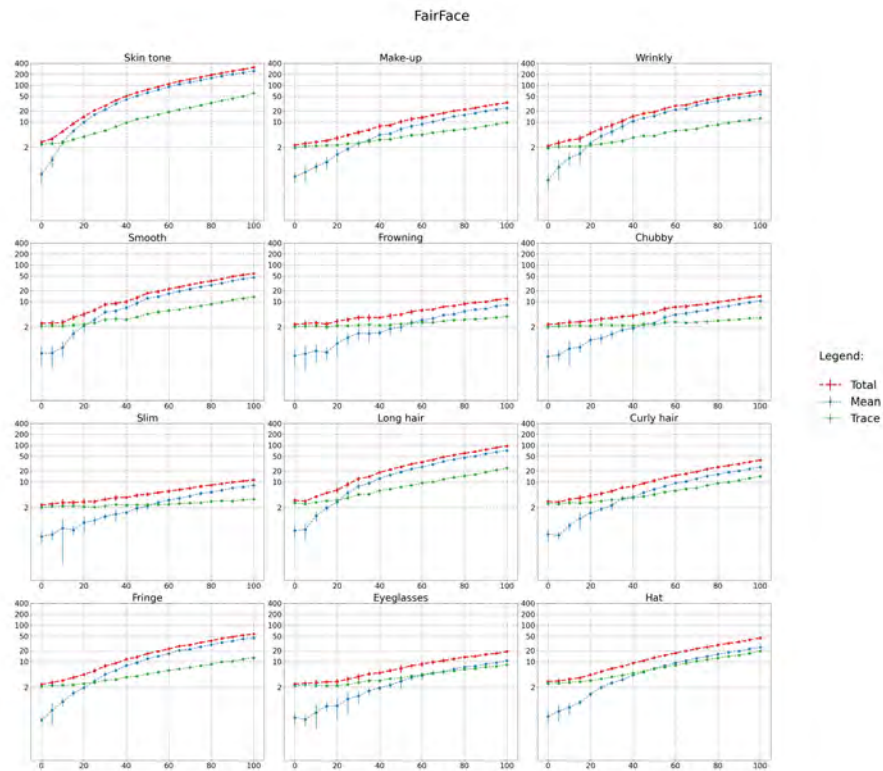


Figure B.4: Results for causal sensitivity analysis of Fréchet distances in the FairFace feature space.

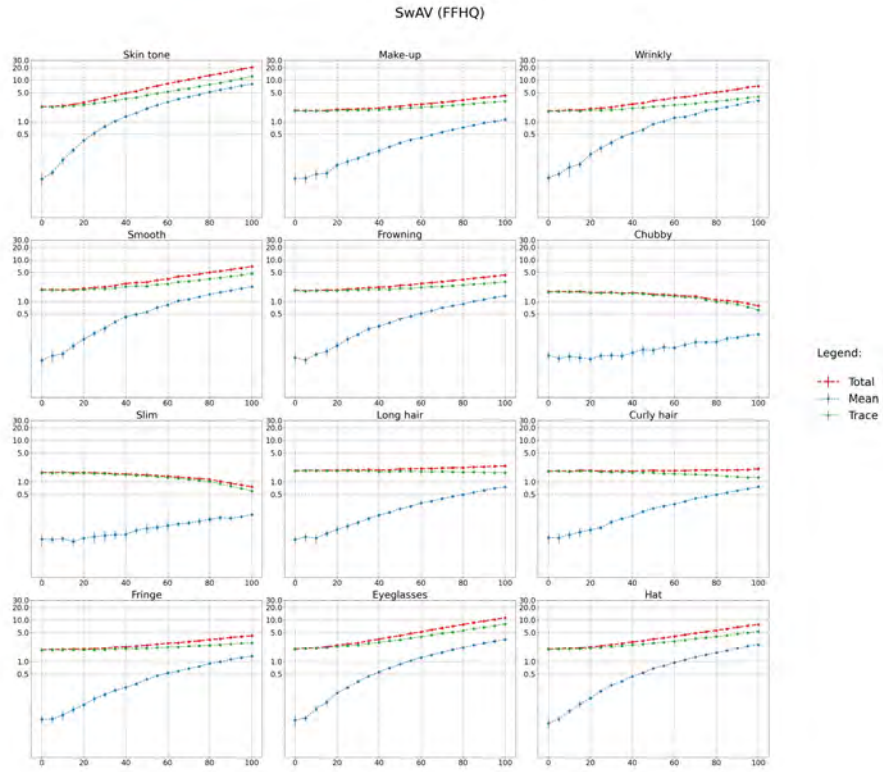


Figure B.5: Results for causal sensitivity analysis of Fréchet distances in the SwAV-FFHQ feature space.

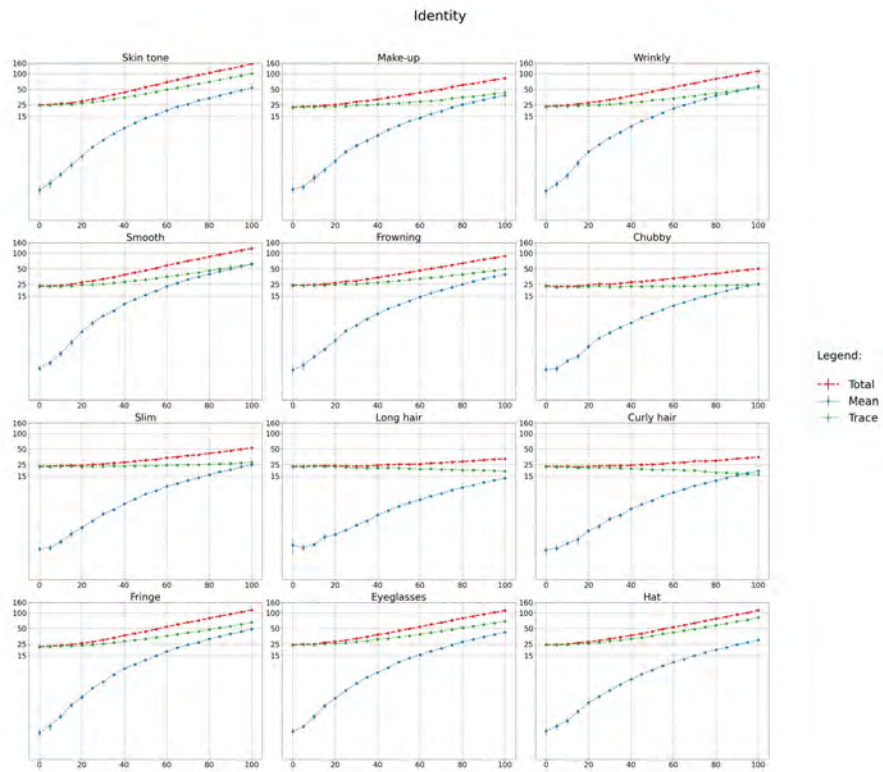


Figure B.6: Results for causal sensitivity analysis of Fréchet distances in the identity feature space.

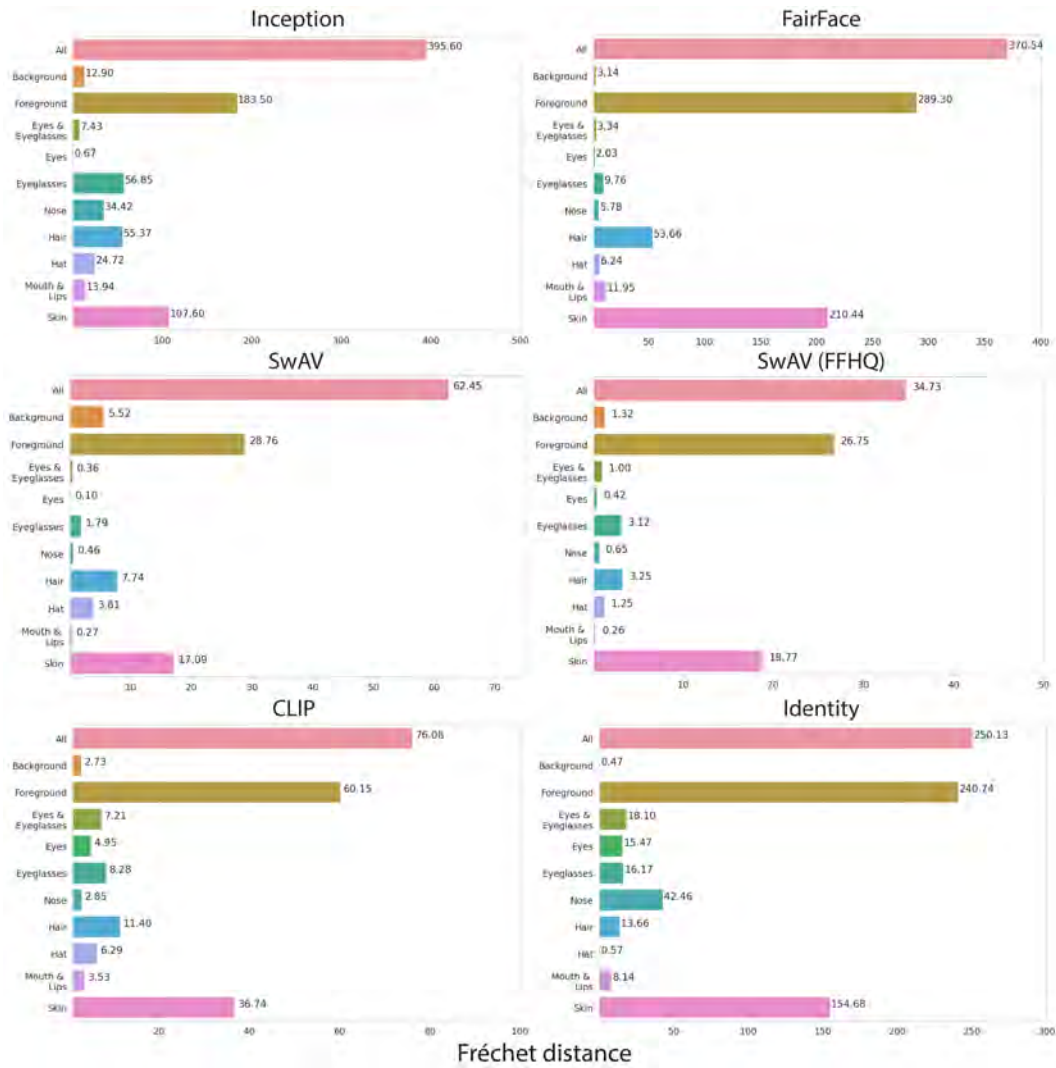


Figure B.7: Results for the effect of semantic region distortion (blur) on Fréchet distances (FD) across different feature spaces.