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

Supplementary Material

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 S1 summarizes the models, code repositories, and checkpoint links used in our implementation.

Table S1. 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!AswpsDO2toNKq0lWY69vN58GR6mw?e=p9Ov5d

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,

Table S2. 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" Target: "face with eyeglasses"	$\alpha = 10$ $\beta = 0.13$
Skin tone	OLLT	N/A	Step size $= 0.5$
Make-up	StyleCLIP	Neutral: "face" Target: "face with makeup"	$\alpha = 3$ $\beta = 0.12$
Wrinkly	StyleCLIP	Neutral: "face with skin" Target: "face with wrinkly skin"	$\alpha = 3$ $\beta = 0.09$
Smooth	StyleCLIP	Neutral: "face with skin" Target: "face with wrinkly skin"	$\alpha = -3$ $\beta = 0.09$
Chubby	StyleCLIP	Neutral: "face" Target: "chubby face"	$\alpha = 5$ $\beta = 0.25$
Slim	StyleCLIP	Neutral: "face" Target: "chubby face"	$\alpha = -5$ $\beta = 0.25$
Frowning	StyleCLIP	Neutral: "smiling face" Target: "frowning face"	$\alpha = 5$ $\beta = 0.20$
Hair length	StyleCLIP	Neutral: "face with hair" Target: 'face with long hair"	$\alpha = 15$ $\beta = 0.20$
Curly	StyleCLIP	Neutral: "face with hair" Target: 'face with curly hair"	$\alpha = 5$ $\beta = 0.25$
Fringe	StyleCLIP	Neutral: "face with hair" Target: 'face with fringe hair"	$\alpha = 5$ $\beta = 0.15$

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 hair[1]. 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[2] analyzed in our experiments (see first column of Table S2). To accomplish this, we utilize one of three different image manipulation methods based on the attribute type: (1) OLLT, (2) StyleCLIP [7], and (3) image inpainting with Stable Diffusion [10]. 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 attribute is listed in Table S2. 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 [5]. 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 S2. 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 S1). 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.

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

[2] We include 4 additional attributes (make-up, slim, curly, fringe) in our supplementary analysis, which were omitted from the main paper due to brevity.

B. Additional experimental results

B.1. Breakdown of FD mean and trace terms

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 S1 to S6 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 S7 plots the (unnormalized) FD for each feature space when the specified semantic region is heavily blurred.

B.2. Analysis of face generators in different feature spaces

In this section, we present evaluations of four popular, publicly available face generation models using metrics computed in each feature space: StyleGAN2 [3], EG3D [2], latent diffusion model (LDM) [8], and Nouveau variational autoencoder (NVAE) [9]. For StyleGAN2 and EG3D, we evaluate the models both with and without truncation [1, 6] ($\psi = 0.7$, truncation cutoff $= 14$). We evaluate models using FD and $k-$nearest neighbors precision and recall metrics [4]. 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 S3.

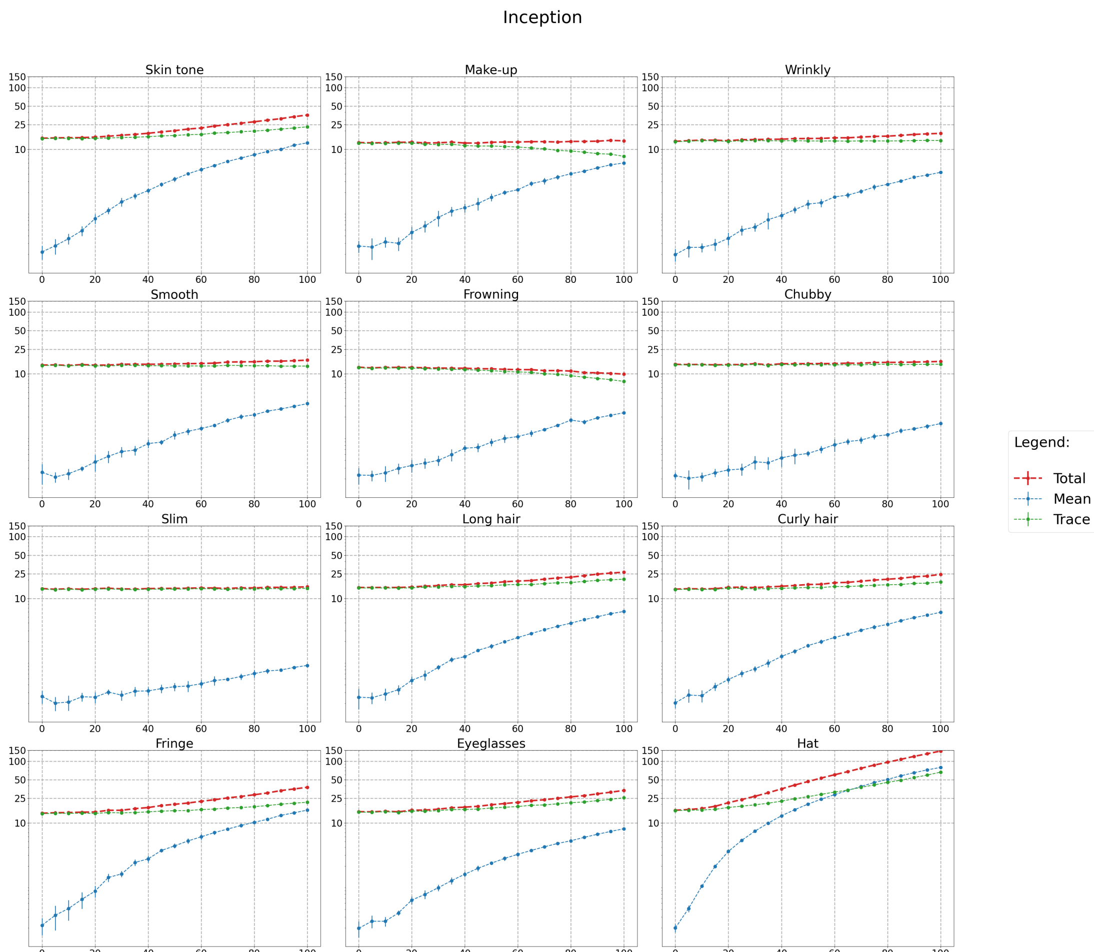

Figure S1. Results for causal sensitivity analysis of Fréchet distances in the Inception feature space.

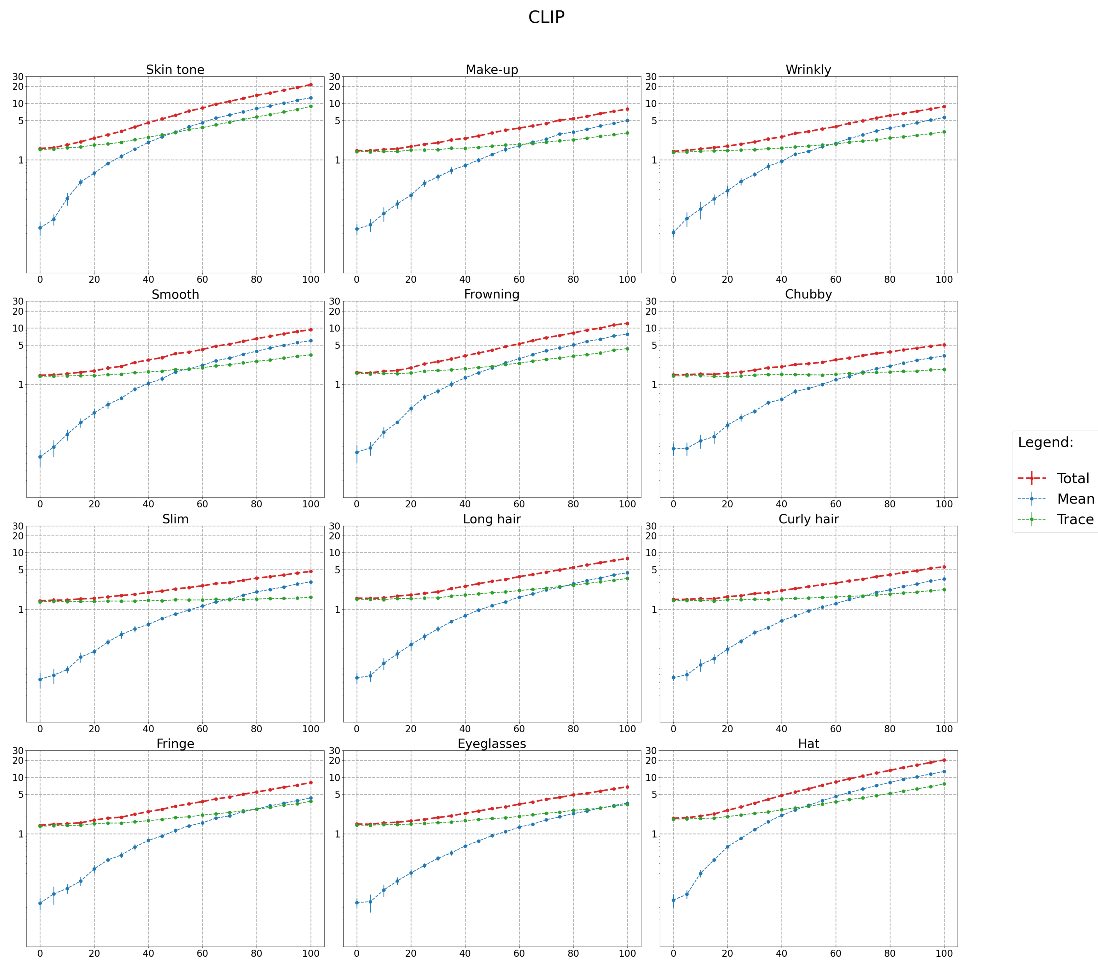

Figure S2. Results for causal sensitivity analysis of Fréchet distances in the CLIP feature space.

References

[1] Andrew Brock, Jeff Donahue, and Karen Simonyan. Large scale GAN training for high fidelity natural image synthesis. In *International Conference on Learning Representations*, 2019. 3

[2] 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. 3, 10

[3] 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. 3, 10

[4] 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*. Curran Associates, Inc., 2019. 3

[5] 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, 2019. 2

[6] Marco Marchesi. Megapixel size image creation using generative adversarial networks, 2017. 3

[7] 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. 2

[8] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-Resolution Image Synthesis With

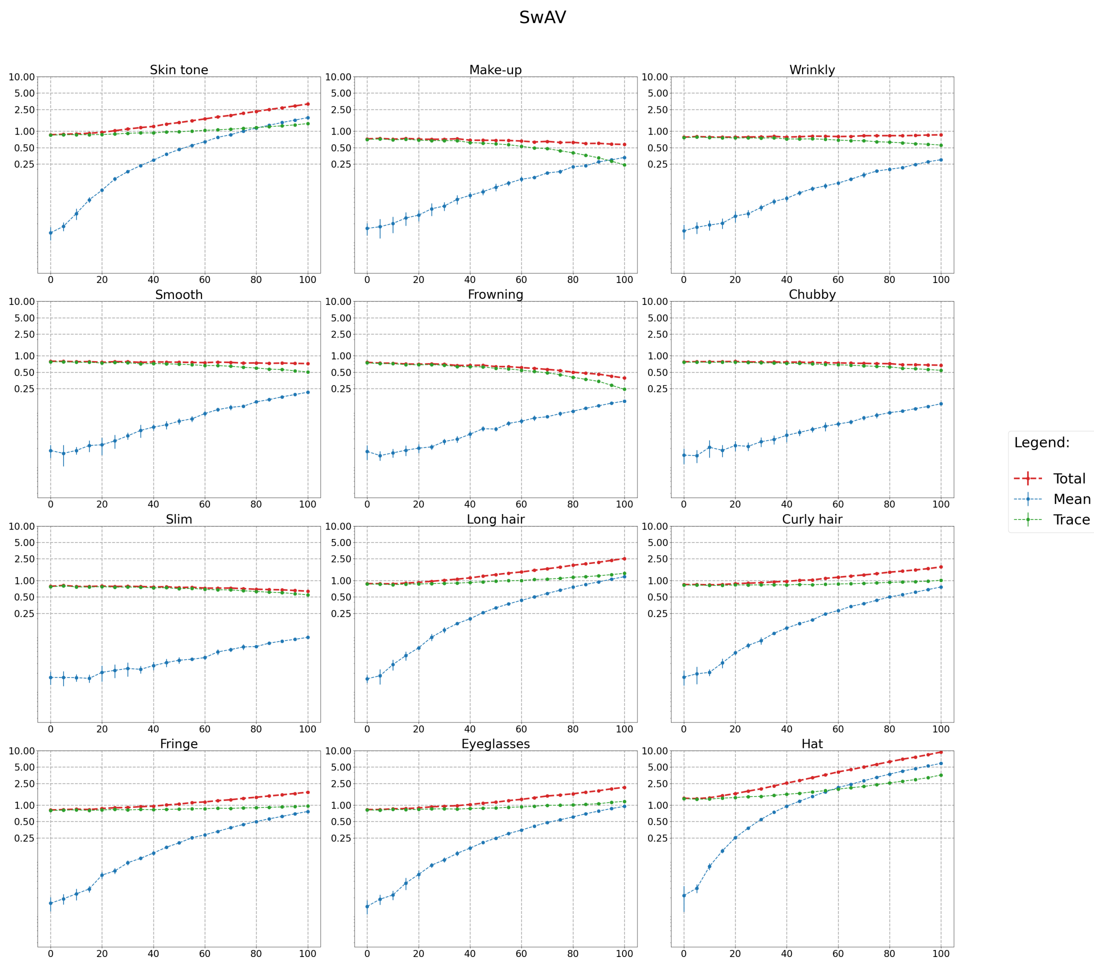

Figure S3. Results for causal sensitivity analysis of Fréchet distances in the SwAV feature space.

Latent Diffusion Models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10684–10695, 2022. 3, 10

[9] Arash Vahdat and Jan Kautz. NVAE: A Deep Hierarchical Variational Autoencoder. In *Advances in Neural Information Processing Systems*, pages 19667–19679. Curran Associates, Inc., 2020. 3, 10

[10] 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, 2023. 2

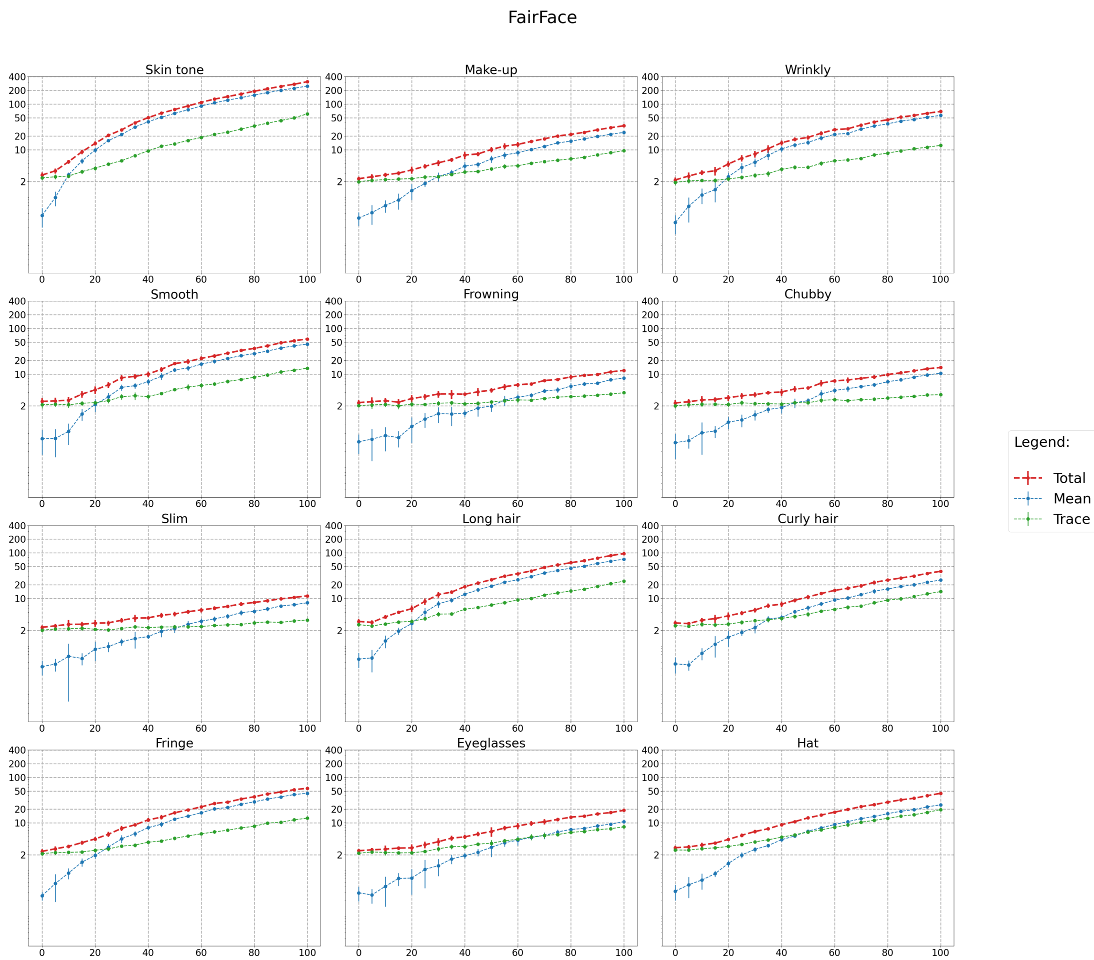

Figure S4. Results for causal sensitivity analysis of Fréchet distances in the FairFace feature space.

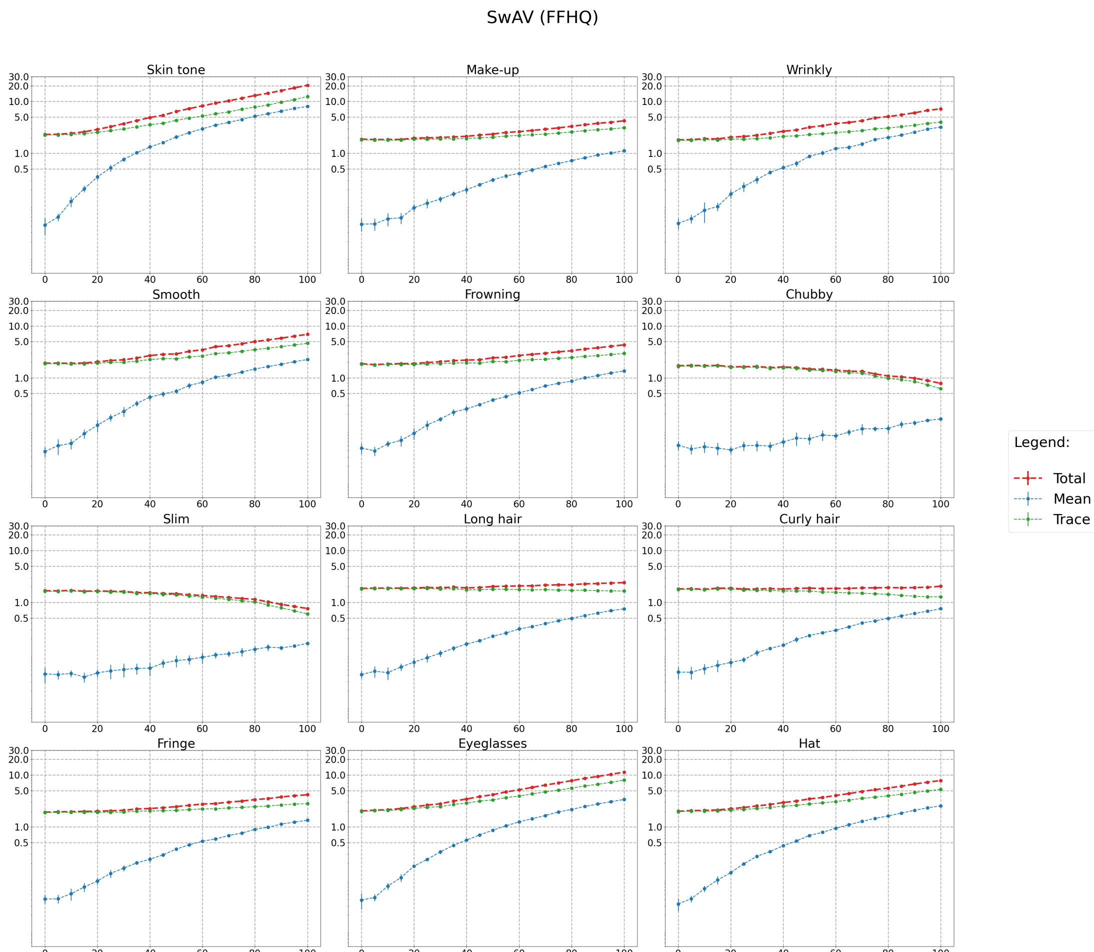

Figure S5. Results for causal sensitivity analysis of Fréchet distances in the SwAV-FFHQ feature space.

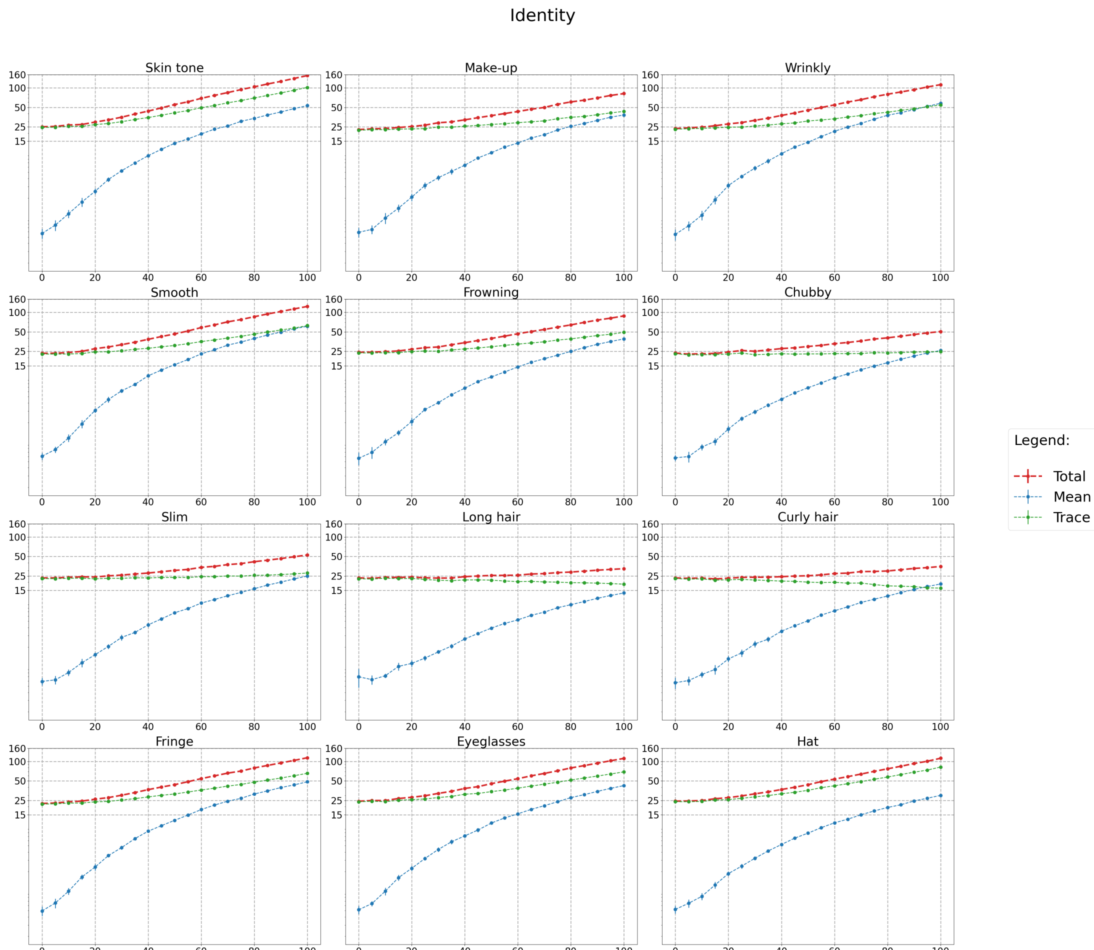

Figure S6. Results for causal sensitivity analysis of Fréchet distances in the identity feature space.

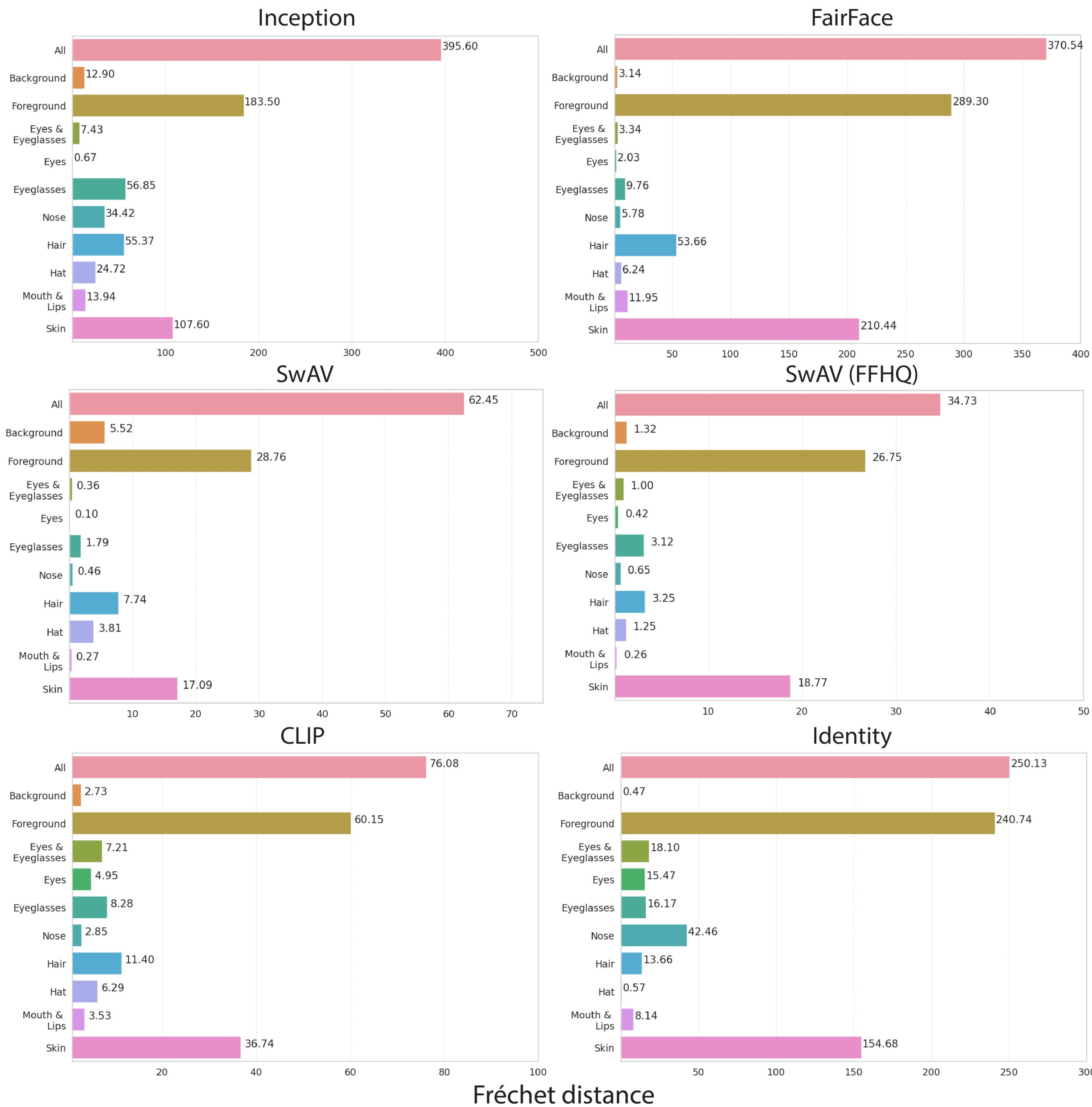

Figure S7. Results for the effect of semantic region distortion (blur) on Fréchet distances (FD) across different feature spaces.

Table S3. **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 (↓)

	Inception	CLIP	SwAV	FairFace	SwAV (FFHQ)	Identity
StyleGAN2 [3]	**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 [2]	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 [8]	**10.0**	**3.6**	**1.7**	**10.9**	**1.4**	**6.9**
NVAE [9]	35.9	9.7	5.4	56.9	5.8	44.1

Precision (%) (↑)

	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 (%) (↑)

	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