All Roads Lead to Rome? Exploring Representational Similarities Between Latent Spaces of Generative Image Models

Charumathi Badrinath CHARUMATHIBADRINATH@COLLEGE.HARVARD.EDU
Usha Bhalla USHA_BHALLA@G.HARVARD.EDU
Alex Oesterling AOESTERLING@G.HARVARD.EDU
Suraj Srinivas SSRINIVAS@G.HARVARD.EDU
Himabindu Lakkaraju HLAKKARAJU@HBS.EDU
Harvard University

Abstract

Do different generative image models secretly learn similar underlying representations? We investigate this by measuring the latent space similarity of four different models: VAEs, GANs, Normalizing Flows (NFs), and Diffusion Models (DMs). Our methodology involves training linear maps between frozen latent spaces to "stitch" arbitrary pairs of encoders and decoders and measuring output-based and probe-based metrics on the resulting "stitched" models. Our main findings are that linear maps between latent spaces of performant models preserve most visual information even when latent sizes differ; for CelebA models, *gender* is the most similarly represented probe-able attribute. Finally we show on an NF that latent space representations converge early in training.

1. Introduction

Recent literature on deep networks has suggested that all models are converging to the same representation of "reality". Huh et al. [6] show that the embedding spaces of performant vision and language models learn similar notions of "distance" between datapoints of the appropriate modalities. Bansal et al. [2] "stitch" together layers of pairs of powerful vision networks via a simple map and show that the combined network has minimal loss in performance.

Asperti and Tonelli [1] investigate whether a similar conclusion holds for *latent spaces* of generative image models by extending the model stitching framework. They train a *linear* map between the latent spaces of a Variational Autoencoder (VAE) and Generative Adversarial Network (GAN) and show that this stitched VAE-encoder-GAN-decoder model is able to produce reconstructions of images that look similar to the originals and have a low RMSE.

However, despite previous work, it remains unclear how well these conclusions hold for other families of generative image models, especially those with vastly different latent sizes. Further, it is not clear which *specific* visual attributes are represented similarly by models and which aren't. In our work, we address these drawbacks by (1) extending the study to include NFs and DMs, (2) including probe-based metrics to capture semantics of representations. On CelebA we find that powerful models can be stitched via their latent spaces to reconstruct an image with little loss in quality. In addition to color, pose, and lighting, attributes related to *gender* are most similarly represented. Finally, we explore the question of how quickly the local minimum corresponding to this shared latent space is found, by seeing how early in training latent space probe accuracy plateaus. On an NF, we find that this happens less than 20% of the way into training.

Our overall contributions are:
(1) In §3.2, we propose two sets of metrics (reconstruction-based, probe-based) grounded in a model "stitching"-based procedure to measure the similarity of generative image model latent spaces.
(2) In §4, we perform extensive experiments across four generative image model classes trained on CelebA (VAEs, GANs, NFs and DMs), and find significant similarity among their latent spaces on our metrics, especially in their representation of gender. On the Normalizing Flow model, we find that this common latent space structure is learned early in training.

2. Related Works

Exploring latent spaces. Several supervised [15] [24] and unsupervised [7] approaches have been proposed to discover vectors corresponding to high-level image concepts (e.g. hair color) in the GAN latent space. These techniques have been applied to find edit directions in the latent spaces of VAEs [29] and NFs [11] as well. Our work *compares* the structure of different models' latent spaces and is thus orthogonal to the exploration of a model's latent space in isolation.

Metrics to compare representations. CCA [19] and CKA [13] are metrics that have been used to measure representational similarity between layers of different models. Another work compared distances between sets of datapoints in different models' embedding spaces [6].

Model stitching to compare representations. One branch of work compares representations via "stitching" where a simple map is learned between layers of two different models [16]. One work stitched together neural networks of different strengths at an intermediate layer, concluding that good models tended to converge on similar representations and that stitching strong models to weak models improved their performance [2]. Work on stitching the latent spaces of generative models has also been explored. One paper showed that it is possible to construct a *linear* map between the latent space of image encoders and text decoders to obtain text describing an image, demonstrating that models learn similar representations of multi-modal data [18].

The work most closely related to ours showed that it is possible to linearly stitch the latent spaces of a VAE and GAN and reconstruct an image with little loss in quality [1]. Our work extends this idea to more architectures; we also identify what specific concepts models represent similarly.

3. Methodology

3.1. Models and Datasets

We experiment on 5 generative image models trained on CelebA [17] – a dataset of celebrity faces annotated with 40 binary attributes. The models and their corresponding latent space sizes are: GAN (512), VAE (512), VQVAE (768), NF (12288), DM (12288). Models output images of size 64×64 with slightly different crops.

We use a GAN-generated dataset *CelebA-Synthetic* to train linear maps to and from the GAN latent space, since GANs lack an encoder. The "latent space" of the DM is taken to be the noised image after running the forward process using the DDIM scheduler with 50 timesteps [25]. As a baseline, we use a "random" encoder which maps images deterministically to arbitrary 512-dimensional standard Gaussian vectors.

3.2. Proposed Metrics

Model latent spaces are first stitched via linear maps (see Appendix A.4).

Reconstruction-Based Metrics. We qualitatively assess the image reconstructions of stitched models. We also compute the latent space MSE of the mapping, pixel-space RMSE of the reconstructed image, LPIPS score [30] which measures the perceptual similarity between a pair of images, and the FID [4] which measures how similar the distribution of reconstructed images is to the distribution of target images. Lower is better for all metrics.

Probe-Based Metrics. We train linear probes (see Appendix A.5) on the latent spaces of encoder-decoder models to detect the presence of binary attributes and report test set accuracy. We then measure the percentage of times probe predictions on latents encoded by the model whose latent space it was trained on and those mapped from other latent spaces match (higher is better). We also measure the percent change in probe accuracy on these sets of latents (near 0 is better).

4. Results

4.1. Assessing Stitched Model Reconstructions

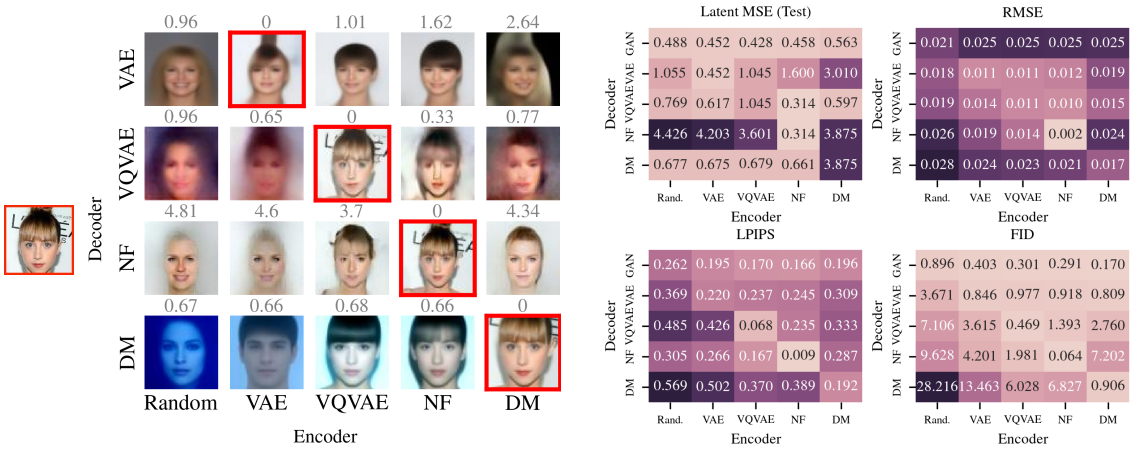

Figure 1: **Left.** Reconstruction of a CelebA image using stitched models. The leftmost image is the ground truth. **Right.** Heatmaps of latent-space MSE, pixel-space RMSE, LPIPS and FID for images reconstructed by stitched models. Stitched models using the VQVAE and NF encoders yield the best reconstructions and lowest metric values.

We stitch the latent spaces of a VAE, VQVAE, NF, DM and GAN pairwise and compute the reconstruction-based metrics described in §3.2. From Figure 1 we see that stitching the NF and VQVAE encoders to any model $\mathbb{X}$'s decoder yields reconstructions resembling those generated by

Figure 2: Reconstruction of CelebA-Synthetic image by stitching to GAN decoder. VQVAE, NF encoders yield the closest reconstructions.

$\mathbb{X}$ (boxed in red) and obtain relatively low latent MSE, RMSE, LPIPS and FID scores. Additional examples are in Appendix A.7.

Conversely, the DM and VAE encoders barely outperform the random encoder in terms of stitched reconstruction quality. Poor VAE performance can be attributed to significant information loss during encoding, reflected in lossy reconstructions even with its own decoder. For the DM, we note that the "latent vectors" we use to construct our mappings are heavily noised versions of the image rather than true encodings, resuling in poor linear probe performance (Figure 3). Latent space maps pick up on remaining signal, which is sometimes enough to map to an area of the decoder model's latent space producing images in the visual vicinity of an target image (Figure 2). The spatial nature of DM "latents" (the"representation" of a feature is a function of the pixel values in the region of the image that the feature manifests) makes the ability to stitch another model's encoder to its decoder *at all* unexpected. However, it explains why models stitched to the DM decoder produce reconstructions preserving the geometry of the target image but not the colors.

Finally, Figure 2 shows reconstructions of a test image from CelebA-Synthetic obtained by stitching various models' encoders to the GAN decoder. Due to the strength of the GAN decoder, all reconstructions resemble a face. Furthermore, mapping to the vicinity of the true GAN latent is sufficient to produce a reconstruction preserving most characteristics of the target image (e.g. age, hair and skin color, expression) supporting results from literature [1].

4.2. Identifying Similarly Represented Attributes

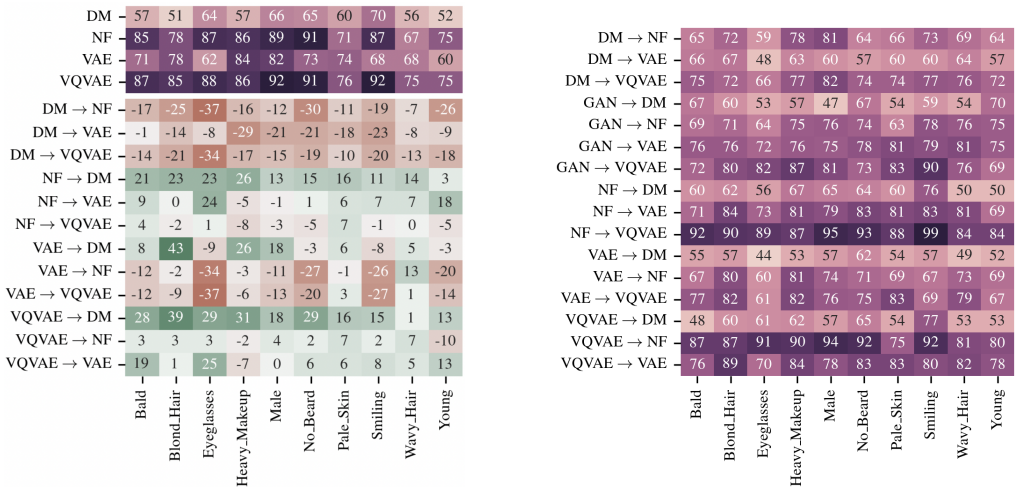

Figure 3: **Left.** Top 4 rows show latent space probe accuracies. NF and VQVAE have the most linear representations. Remaining rows show change in accuracy of probe trained on model $\mathbb{X}$'s latent space when making predictions on latents encoded by $\mathbb{X}$ and latents mapped from $\mathbb{X}' \to \mathbb{X}$. **Right.** Percentage of times a probe trained on model $\mathbb{X}$'s latent space produces the same prediction on latents encoded by $\mathbb{X}$ and latents mapped from $\mathbb{X}' \to \mathbb{X}$. Attributes correlated with gender are represented similarly by nearly every model.

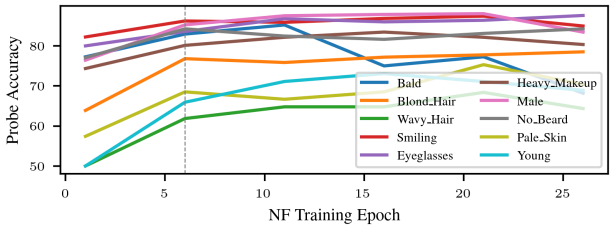

Figure 4: Test accuracy of probes on NF latent space frozen at increasing training epochs; accuracy plateaus after epoch 6.

We stitch the latent spaces of a VAE, VQVAE, NF, DM and GAN pairwise and compute the probe-based metrics described in §3.2 on 10 interesting CelebA attributes. Results for all 40 attributes are in Appendix A.6. From Figure 3 we see that linear probes for the gender-related attributes `Male` and `Heavy Makeup` achieve 80% accuracy on the NF, VAE and VQVAE latent spaces. In Figure 3 we see that NF probes produce similar predictions on latents produced by its own decoder and latents encoded and mapped from the VQVAE latent space (and vice versa), in agreement with the results from §4.1.

We hypothesize that attributes that are represented similarly in all models' latent spaces are either uniformly represented in the training dataset (e.g. `Smiling`), or explain much of the variance in an image (e.g. `Male` and related attributes). Interestingly, post-map probe match percentages are high for gender-related attributes, particularly `Heavy_Makeup` and `Male`, even for for maps originating from the non-linear DM latent space indicating that signal corresponding to these attributes is least likely to be lost in the noising process.

Finally, we note in Figure 3 that probes trained on less linearly separable latent spaces are often *more* accurate when applied on latents mapped from a more linearly separable latent space, explaining why corresponding reconstructions have more "extreme" versions of the attributes present in the target image (e.g. distinct, block-like bangs in NF $\rightarrow$ DM and VQVAE $\rightarrow$ DM maps in Figure 1).

4.3. Observing Latent Space Structure Over Training

We train the Glow NF [11] on CelebA for 26 epochs checkpointing every 5 epochs. We train linear probes for 10 interesting attributes on the frozen latent spaces and measure how their accuracy changes as model training progresses. Figure 4 shows that probe accuracy begins to plateau for all attributes by epoch 6 of 26. Attributes correlated with gender as well as `Smiling` (the similarly represented attributes identified in §4.2) change minimally in their representations after epoch 1 of 26. Thus, we hypothesize that a "universal" representation is natural for models to learn and that later training epochs improve the decoder given a somewhat fixed latent representation.

5. Discussion and Future Work

This work advances the hypothesis that as (unconditional) generative image models become more expressive, their latent space representations (given that they are meaningful) become more similar to one another. We find that this shared representation is learned early on in model training, indicating that it is quite "natural" in describing the data. This suggests that the latent spaces of increasingly expressive models may be converging to an "optimal" data representation in the limit, which has deep implications for how we construct world models.

Similarity of latent space representations also means that edit directions, unsafe regions, or areas of bias found in the latent space of one model, can be mapped onto any other model via a simple

linear map. It also allows for more powerful image editing by stitching models with strong encoders to models with strong decoders via their latent spaces.

Future work can investigate whether similar results hold for models trained on datasets containing images from multiple classes with more variance, and can probe similarity of representation of non-binary attributes. We expect to see similar results given that the models used are able to produce realistic samples from the data distribution. We also wish to extend our analysis to pipelines composed of *multiple* foundation models, for example, the latent diffusion model [23] which trains a DM on the VQVAE latent space. Since it is the *first* foundation model in this pipeline that learns a latent representation of the data, we expect to see similar results to this paper. Finally, work has shown that it is possible to find a linearly separable latent space for DMs [14]; we can validate that this latent space is similarly structured to those studied in our paper.

Acknowledgements

We would like to thank the GRaM workshop's ELLIS Mobility Grant as well as the SPIGM workshop for supporting CB's travel and attendance at ICML.

References

[1] Andrea Asperti and Valerio Tonelli. Comparing the latent space of generative models, 2022.

[2] Yamini Bansal, Preetum Nakkiran, and Boaz Barak. Revisiting model stitching to compare neural representations, 2021.

[3] Ian J. Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks, 2014.

[4] 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, 2018.

[5] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models, 2020.

[6] Minyoung Huh, Brian Cheung, Tongzhou Wang, and Phillip Isola. The platonic representation hypothesis, 2024.

[7] Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. Ganspace: Discovering interpretable gan controls, 2020.

[8] Tero Karras, Samuli Laine, and Timo Aila. A style-based generator architecture for generative adversarial networks, 2019.

[9] Tero Karras, Miika Aittala, Janne Hellsten, Samuli Laine, Jaakko Lehtinen, and Timo Aila. Training generative adversarial networks with limited data, 2020.

[10] Seonghyeon Kim. glow-pytorch. https://github.com/rosinality/glow-pytorch, 2020.

[11] Diederik P. Kingma and Prafulla Dhariwal. Glow: Generative flow with invertible 1x1 convolutions, 2018.

[12] Diederik P Kingma and Max Welling. Auto-encoding variational bayes, 2022.

[13] Simon Kornblith, Mohammad Norouzi, Honglak Lee, and Geoffrey Hinton. Similarity of neural network representations revisited, 2019.

[14] Mingi Kwon, Jaeseok Jeong, and Youngjung Uh. Diffusion models already have a semantic latent space, 2023.

[15] Anders Boesen Lindbo Larsen, Søren Kaae Sønderby, Hugo Larochelle, and Ole Winther. Autoencoding beyond pixels using a learned similarity metric, 2016.

[16] Karel Lenc and Andrea Vedaldi. Understanding image representations by measuring their equivariance and equivalence, 2015.

[17] Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*, December 2015.

[18] Jack Merullo, Louis Castricato, Carsten Eickhoff, and Ellie Pavlick. Linearly mapping from image to text space, 2023.

[19] Ari S. Morcos, Maithra Raghu, and Samy Bengio. Insights on representational similarity in neural networks with canonical correlation, 2018.

[20] Kushagra Pandey, Avideep Mukherjee, Piyush Rai, and Abhishek Kumar. VAEs meet diffusion models: Efficient and high-fidelity generation. In *NeurIPS 2021 Workshop on Deep Generative Models and Downstream Applications*, 2021. URL https://openreview.net/forum?id=-J8dM4ed_92.

[21] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[22] Danilo Jimenez Rezende and Shakir Mohamed. Variational inference with normalizing flows, 2016.

[23] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2022.

[24] Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. Interpreting the latent space of gans for semantic face editing, 2020.

[25] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv:2010.02502*, October 2020. URL https://arxiv.org/abs/2010.02502.

[26] Aaron van den Oord, Oriol Vinyals, and Koray Kavukcuoglu. Neural discrete representation learning, 2018.

[27] Patrick von Platen, Suraj Patil, Anton Lozhkov, Pedro Cuenca, Nathan Lambert, Kashif Rasul, Mishig Davaadorj, Dhruv Nair, Sayak Paul, Steven Liu, William Berman, Yiyi Xu, and Thomas Wolf. Diffusers: State-of-the-art diffusion models. URL https://github.com/huggingface/diffusers.

[28] Zhendong Wang, Huangjie Zheng, Pengcheng He, Weizhu Chen, and Mingyuan Zhou. Diffusion-gan: Training gans with diffusion. *arXiv preprint arXiv:2206.02262*, 2022. URL https://arxiv.org/abs/2206.02262.

[29] Tom White. Sampling generative networks, 2016.

[30] Richard Zhang, Phillip Isola, Alexei A. Efros, Eli Shechtman, and Oliver Wang. The unreasonable effectiveness of deep features as a perceptual metric, 2018.

Appendix A.

A.1. Source Code

All code for this paper is available in the following GitHub repository:
https://github.com/charumathib/thesis-latent-spaces.

A.2. Background on Models

All generative image models learn to "sample" images following their the distribution of their training set. However, the four main families of models – Variational Autoencoders (VAEs), Generative Adversarial Networks (GANs), Normalizing Flows (NFs) and Diffusion Models (DMs) – differ drastically in how they do so.

A **VAE** [12] consists of an encoder network mapping images to latent vectors of lower dimension, and a decoder neural network that maps latent vectors to images with the objective of minimizing image reconstruction error. A KL-Divergence term is added to the objective to encourage a marginally Gaussian latent space distribution.

A **GAN** [3] is comprised of a generator network (decoder) that learns to generate images from marginally Gaussian latent samples and a discriminator network whose objective is to distinguish images from the training set from generated images. Both networks improve by playing a minimax game, ultimately yielding a generator capable of producing realistic images.

An **NF** [22] samples a latent variable from a simple distribution and repeatedly applies a series of invertible functions to this variable until it is transformed into an image. The marginal likelihood of each datapoint can be derived via the change of variables theorem for distributions and the objective is the negative data log likelihood.

Finally, a **DM** [5] progressively adds Gaussian noise to an image producing a sequence noisy samples that become isotropic to a Gaussian in the limit ("forward process"). It then learns how to produce samples from the "reverse process" to gradually transform pure noise into an image. DMs dont have an obvious latent space in the same way as VAEs, GANs and NFs since marginally Gaussian "latents" are created via progressive noising rather then by a transformation that would force a representation to be learned.

A.3. Models Used

The pre-trained GAN we use is from the Diffusion-GAN codebase [28] and follows the StyleGAN2-ADA architecture [9]. The pre-trained VAE is from the DiffuseVAE codebase [20]. The pre-trained VQVAE [26] from the Hugging Face `diffusers` library [27]. The NF implementation is from an independent source [10] follows the Glow architecture [11]. The DM is from the DDIM codebase [25].

A.4. Training Latent Space Linear Maps

Many of our experiments rely on the construction of linear maps between model latent spaces. For each pair of encoder-decoder models (VAE, VQVAE, NF, DM), we encode the first 9000 images from the CelebA test split into their respective latent spaces and use a closed-form linear regression solver from `sklearn` to predict one set of latents from the other [21]. To prevent overfitting in maps originating from the NF and DM latent spaces, we use ridge regression with manually tuned values of α as summarized in Table 1.

Map	α
DM $\rightarrow$ GAN	2000
DM $\rightarrow$ VAE	100
DM $\rightarrow$ VQVAE	5000
DM $\rightarrow$ NF	5000
NF $\rightarrow$ GAN	50000
NF $\rightarrow$ VAE	5000
NF $\rightarrow$ VQVAE	50000
NF $\rightarrow$ DM	50000

Table 1: Regularization parameters used in Ridge regression for each linear map. Maps that are not listed in this table were not regularized.

Our training procedure differs slightly for maps to and from the GAN latent space. We use the "style space" as our latent space (a layer connected to the base latent space by an 8 layer MLP) [8]. We use the first 9000 images in the CelebA-Synthetic dataset to train maps in the way described earlier; the latent used by the GAN to generate each image is treated as the image's "encoding" in the GAN latent space. All maps are tested on a holdout set of 100 images from either CelebA or CelebA-Synthetic as appropriate.

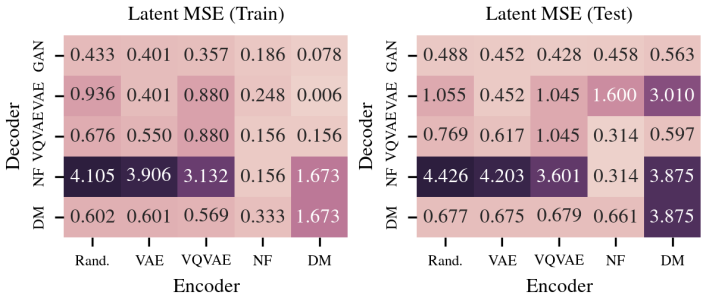

Figure 5: Latent space map MSE on train (left) and test (right) sets.

A.5. Training Latent Space Linear Probes

We train linear probes to predict the value of each of the 40 binary attributes (e.g. `Male`) CelebA images are annotated with. To construct a balanced training set for an attribute's probe, we first compute the number of images with and without that attribute in the first 9000 CelebA images. 80% of the minimum of these two numbers is how many positive and negative examples are included. Probes are trained using lasso regression in `sklearn` where $\alpha = 0.005$ for probes on the VAE and VQVAE latent spaces, 0.02 for probes on the DM latent space and 0.1 for probes on the NF latent space. [21]. We use a holdout set of 200 latents (100 with attribute, 100 without attribute) to evaluate probe performance.

A.6. Full Probe Results

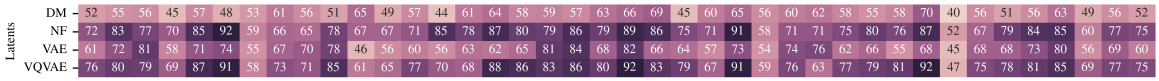

Figure 6: Full version of Figure 3 (left, top) with results for all 40 attributes.

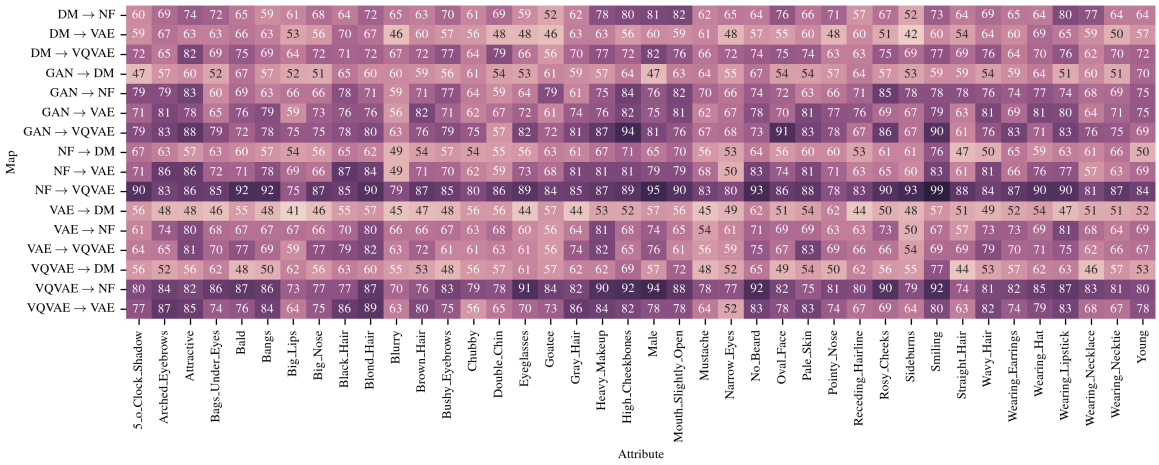

Figure 7: Full version of Figure 3 (left, bottom) with results for all 40 attributes.

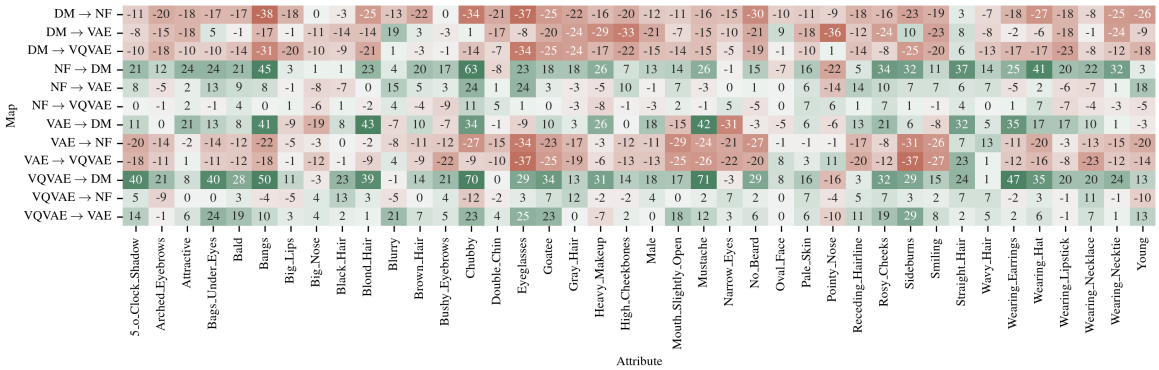

Figure 8: Full version of Figure 3 (right) with results for all 40 attributes.

A.7. Additional Examples of Stitched Model Reconstructions

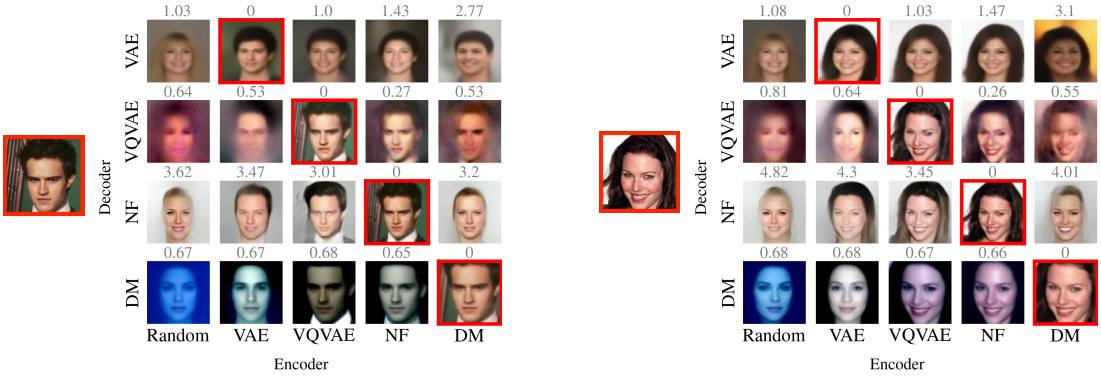

Figure 9: See caption of Figure 1.

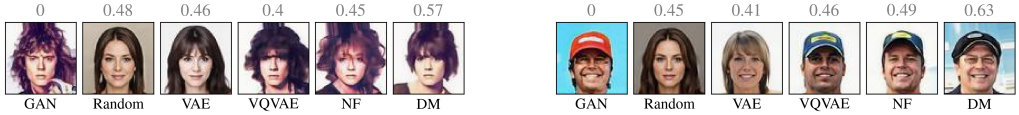

Figure 10: See caption of Figure 2.