This CVPR Workshop paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the accepted version;
the final published version of the proceedings is available on IEEE Xplore.

Language-Guided Trajectory Traversal in Disentangled Stable Diffusion Latent Space for Factorized Medical Image Generation

Zahra TehraniNasab* McGill University
Amar Kumar* McGill University
Tal Arbel McGill University

zahra.tehraninasab@mail.mcgill.ca, amar.kumar@mail.mcgill.ca, tal.arbel@mcgill.ca

Abstract

Text-to-image diffusion models have demonstrated a remarkable ability to generate photorealistic images from natural language prompts. These high-resolution, language-guided synthesized images are essential for the explainability of disease or exploring causal relationships. However, their potential for disentangling and controlling latent factors of variation in specialized domains like medical imaging remains under-explored. In this work, we present the first investigation of the power of pre-trained vision-language foundation models, once fine-tuned on medical image datasets, to perform latent disentanglement for factorized medical image generation and interpolation. Through extensive experiments on chest X-ray and skin datasets, we illustrate that fine-tuned, language-guided Stable Diffusion inherently learns to factorize key attributes for image generation, such as the patient's anatomical structures or disease diagnostic features. We devise a framework to identify, isolate, and manipulate key attributes through latent space trajectory traversal of generative models, facilitating precise control over medical image synthesis.

1. Introduction

Deep learning models for medical imaging have shown state-of-the-art performance across several tasks such as disease classification, image segmentation, drug discovery and high-resolution image synthesis. However, these models often struggle to generalize well to new, unseen data due to domain shifts [24] or entanglement in image features [19]. For example, in medical imaging, multiple factors, such as disease pathology and imaging modality, are often intertwined, making it difficult for the model to isolate the relevant features for the specified task. This entanglement can lead to the model relying on spurious correlations, which may not be present in new data, thus hindering its ability to generalize. Thus, disentanglement aims to separate the various factors of variation in the data, allowing the model to learn more robust and interpretable representations. In computer vision, disentanglement models [2, 8, 12] have been shown to help increase generalization [18, 34] and explainability [29] by isolating task-relevant features from confounding factors. This enables the model to adapt to new data distributions. In addition, disentanglement improves model explainability by separating the factors of variation and providing a clearer understanding of how different features contribute to the final prediction. Vision-language Foundation Models (VLFM) have emerged as a powerful approach for learning disentangled representations, offering several advantages over traditional generative architectures due, in large part, to the enormous dataset size they have been trained on. Earlier approaches designed to disentangle latent factors relied on specialized generative architectures such as Variational Auto Encoders (VAE) [15], which don't produce high-resolution images required for medical imaging, or Generative Adversarial Networks (GAN) [4], which are notoriously hard to train. Normalizing Flows [22] have been adapted to address disentanglement in simpler contexts but are prohibitively com-

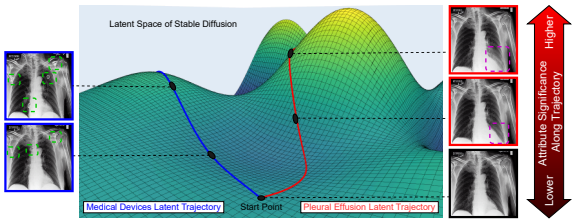

Figure 1. Traversal along the latent trajectories of Stable Diffusion using language guidance. Given an initial chest X-ray projected onto latent space (start point), traversal along the trajectory is performed via language guidance. Sampling along the trajectory results in only a single attribute (e.g. "medical devices", "pleural effusion") being altered from the start point ("neutral"), while the patient identity is maintained.

*Equal contribution

putationally expensive. Other challenges faced when adapting traditional models include their difficulties in training them end-to-end [32, 38], and their requirement for specialized architectural components for conditioning such as AdaIN [9], FiLM [23], SPADE [21]. Furthermore, models often require specific heuristics to permit traversals in latent space [7, 26–28]. In contrast, vision-language foundation models offer more efficient and scalable solutions. The power of these foundation models to disentangle latent representations, enabling targeted image modifications while preserving semantic content and other attributes, has been explored in the natural imaging domain [35, 36]. However, this domain remains under-explored in the context of medical images, where complex entanglement is common and challenging to address.

In this paper, we investigate the disentanglement capabilities of Stable Diffusion [25], fine-tuned on medical images, and propose the first method to traverse latent space based on language guidance and see the factorizing effect on the resulting images. Language guidance is seen to be effective at identifying *attribute-specific trajectories* in the latent space (see Figure 1). Additionally, interpolations between samples permit continuous trajectories that can be sampled. Experiments illustrate that the samples exhibit the same disentangled properties, that is, the attribute of interest remains the same in the image throughout the trajectory, becoming more prevalent proportional to the distance from the starting point. We propose a new metric, *Classifier Flip Rate* along a Trajectory (CFRT), to validate the presence of the desired disentanglement.

2. Methodology

Our method for building disentangled representations follows a two-stage process: (i) fine-tune the pre-trained Stable Diffusion on the relevant dataset (in this paper, CheXpert or ISIC); (ii) denoising diffusion implicit model (DDIM) [30] conditioned on the text embedding to synthesize images along the trajectory (see Figure 2).

2.1. Text-to-Image Generation using fine-tuned Stable Diffusion

Given a text embedding, e, the objective of the text-conditioned fine-tuned DDIM is to generate an image, X. The forward diffusion process starts at the timestep $t = 0$ and noises the image X_0 till timestep $t = T$ resulting in a set of noisy images, $X_{1:T}$. The denoising process of the DDIM attempts to denoise from X_T to the original image X_0 where a noisy image at each step is given by X_{t-1}:

$$X_{t-1} = \gamma_{t_0} X_t + \gamma_{t_1} \epsilon_0 \left(X_t, t, e_t\right). \tag{1}$$

Here, γ_{t_0} and γ_{t_1} are computed as $\gamma_{t_0} = \sqrt{\frac{\alpha_{t-1}}{\alpha_t}}$ and $\gamma_{t_1} = \sqrt{1-\alpha_{t-1}} - \sqrt{\frac{\alpha_{t-1}}{\alpha_t} - \alpha_{t-1}}$ respectively with $\alpha_{1:T}$ as the

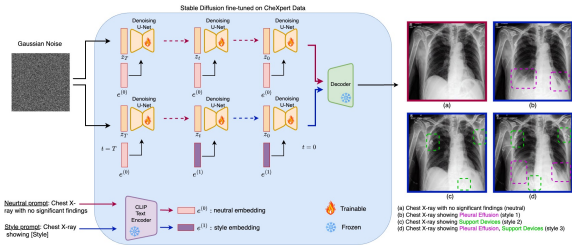

Figure 2. Reverse Diffusion for synthesizing disentangled images from text prompts using fine-tuned Stable Diffusion. The reverse diffusion process takes noisy latents, z_T, as input. A U-Net architecture generates the denoised latent, z_0. During the denoising process, the text embeddings, e_t ($t \in 0, .., T$), from the pre-trained CLIP encoder are added to the latent via cross-attention modules [25]. Finally, the de-noised latent is passed through the decoder to create the synthesised image. Note that the text embeddings can be replaced at some intermediate timestep t during the reverse diffusion.

hyperparameters for the diffusion process. In the usual setting of text-to-image generation, a single text prompt can generate a single image. In this work, image generation depends on the diffusion timestep, t, to explore the properties of the latent representation. Thus, the image generation process, X_0, [35] can be described as in Eq. 2, where $\mathbf{g}$ is the generator:

$$X_0 = \mathbf{g}(X_T, e_{1:T}). \tag{2}$$

Fine-tuning Stable Diffusion (SD): Similar to [13], CheXpert's binary labels are converted into text captions to fine-tune Stable Diffusion (SD) v1.5 [25], which utilizes CLIP embeddings for conditioning. The models are trained using text prompts with the following template: `Chest X-ray of a subject with [finding(s)]` or `a dermoscopic image with [disease] showing [artifacts]` for CheXpert or ISIC respectively (see Section 3.1). To ensure stability while fine-tuning, the CLIP encoder and VAE encoder/decoder remain frozen, and we train the denoising U-NET component. This approach preserves the CLIP encoder's semantic knowledge while adapting the image generation to match the CheXpert data distribution.

Inference: A "neutral" image, X_0 is first generated based on a given text prompt, with embeddings e. To add specific attributes, such as disease(s) or style(s) such as hairs and gel bubbles, to the "neutral" image, a new text prompt with embeddings e' is used. During the reverse diffusion, e' replaces the original text embeddings e at some timestep t. Replacing the embedding at different timestep results in variations in the denoised latent, z_T. The properties of these images are further discussed in section 3.2.

For a given attribute associated with chest X-rays, such as the appearance of support devices (e.g. pacemaker, wires, tubes) or disease states (e.g. pleural effusion), the

disentangled latent representations of the attributes of high-dimensional medical images should follow a non-linear trajectory. With a discrete set of images, a smooth interpolation between samples is made possible through Bézier curves [20], which would be preferable over the more choppy linear interpolation [3].

2.2. Metrics for evaluating the conditionally generated images & attribute disentanglement

We hypothesize that the disentanglement of attributes exists along different trajectories in the latent space of Stable Diffusion, where the latent codes fall along a nonlinear trajectory. We propose a new metric, *Classifier Flip Rate along a Trajectory (CFRT)*, to validate disentanglement along the specified (style) trajectory. In Eq. 3, the CFRT score ensures that an attribute classifier's decision changes more significantly from the starting point ("neutral image") when the attribute of interest, $\mathcal{A}$, is modified compared to changes from modifying any other attribute.

$$\text{CFRT}_{\mathcal{A}} = \frac{1}{|X|}\sum_{x\in X} \mathbb{1}\Big[|f(x) - f(x'_{\mathcal{A}})| > \max_{j\neq\mathcal{A}} |f(x) - f(x'_j)| \wedge\ y(x) = y(x'_{\mathcal{A}}) \wedge\ \forall k \neq \mathcal{A}, x_k = x'_{\mathcal{A}}(k)\Big] \tag{3}$$

where X is the set of all the samples, $x'_{\mathcal{A}}$ is the conditionally generated sample, x is the starting point in the trajectory, where only attribute $\mathcal{A}$ is flipped, $f(.)$ is the classifier's prediction for sample x, $y(x)$ is the ground truth label for sample x, $\mathbb{1}[\cdot]$ is the indicator function with value 1 if the condition is true and 0 otherwise and $\forall j \neq \mathcal{A}, x_j = x'_{\mathcal{A}}(j)$ ensures that all attributes j other than $\mathcal{A}$ remain the same between x and its counterfactual $x'_{\mathcal{A}}$. A higher value of CFRT indicates attribute-specific changes to the image without affecting the other attributes, thus showing disentanglement.

The trajectory can be identified using a text prompt indicating an attribute of interest (e.g. "pleural effusion"). A second metric, the pairwise *cosine similarity* [17] (between the directional vectors), can be used to prove that the latent representations for a particular attribute lie on a non-linear trajectory (thus cannot be captured easily by Style-GAN's [14] singular value decomposition which assumes that the trajectory is linear). More formally, the cosine similarity between the direction of the latent representation of the conditionally generated (z_{cg}) image at timestep t_i (z_{cg,t_i}) and at timestep t_j (z_{cg,t_j}), both relative to the latent of the original ("neutral") image, z_{orig}.

$$\cos(\theta_{t_i,t_j}) = \frac{\Delta\vec{z}_{t_i} \cdot \Delta\vec{z}_{t_j}}{||\Delta\vec{z}_{t_i}|| \cdot ||\Delta\vec{z}_{t_j}||}, \quad \forall t_i, t_j \in T \tag{4}$$

where $\Delta\vec{z}_t = \vec{z}_{\text{cg},t} - \vec{z}_{\text{orig}}$ and $T = \{5k : k \in \{1, 2, ..., 9\}\}$. Finally, the visual quality of the conditionally generated images can be validated using a standard metric, *Learned Perceptual Image Patch Similarity (LPIPS)* [37]. This metric can be computed for the conditionally generated samples, as well as any samples obtained through interpolation along an attribute trajectory relative to the starting point. A lower LPIPS score indicates that two images are more similar/closer in perceptual quality.

3. Experiments & Results

3.1. Dataset and Implementation Details

We perform experiments on two publicly available datasets - CheXpert [10] and ISIC2019 [5, 6, 33]. For CheXpert, `findings` include pleural effusion or support devices while for ISIC2019, `disease` includes melanoma (MEL) and melanocytic nevus (NV). We use the `artifact` attributes, such as hair, gel bubbles, ink and ruler, provided by Bissoto et al. [1][1].

`Stable-Diffusion-v1-5` [25] is used for text-to-image generation (high-resolution, 512×512) tasks. The pre-trained model is frozen, except for the U-net in the denoising process (see Figure 2) [2]. A variant of the DDIM sampler [16] is used for generating images with 50 total backward diffusion steps (for latent samples along the trajectory, we chose the timesteps from $t = 5$ to $t = 45$ in steps of 5). We use 7 sample points to approximate a Bézier curves [20] and sample 50 new latents along this interpolated trajectory to evaluate if these samples exhibit the same properties as their trajectory. For better stability, the variance of the additive Gaussian noise is set to 0 [11].

3.2. Results

Qualitative evaluations: The trajectory in Stable Diffusion latent space corresponding to the desired attribute can be empirically traversed using text representing the target attribute, such as diseases or artifacts. First, we establish the disentanglement capability of text-to-image generation in vision-language foundation models between style (attribute) and content (general anatomy of the patient in the image). Empirically, swapping the initial text prompt (neutral) with the new text prompt (with style) at different time steps in the reverse diffusion process results in variations of the same attribute without changing the content (patient) or other attributes (disease/artifacts) - see Figure 3.

Next, we explore the properties of the latent vectors of the conditionally generated images from Stable Diffusion for different attributes. Empirically, for each specific subject, there exist trajectories in the latent space that cor-

[1] Artifact information for ISIC is available at https://github.com/alceubissoto/artifact-generalization-skin.

[2] https://github.com/tehraninasab/sd-latent-traversal

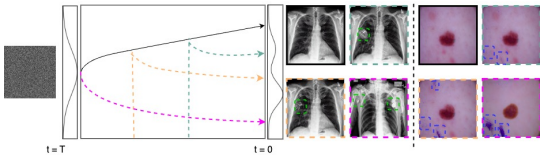

Figure 3. Disentanglement property of the Stable Diffusion [35]. Starting from Gaussian noise (left image) at sampling timepoint t=T, the reverse diffusion process denoises the image (right) at timepoint t=0. The text prompts for the "neutral" images (with **dark borders**) for CheXpert and ISIC are `Chest x-ray with no significant findings` and `A dermoscopic image with melanocytic nevus (NV)`, respectively. The images on the right (matched with coloured borders) are the synthesized images with the same text prompts `Chest x-ray showing Support Devices` for CheXpert and `a dermoscopic image with melanocytic nevus (NV) showing ink` sampled at different timesteps during the reverse diffusion process. Notice that sampling closer to the timepoint t=0 results in a synthesized image similar to the **original image** and as we sample closer to the timepoints t=T, the patient's anatomical structure changes.

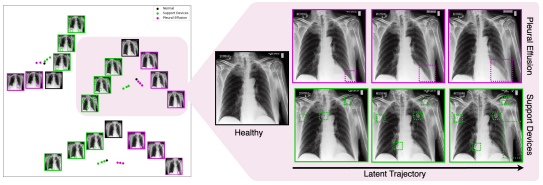

Figure 4. t-SNE plot of generated latent vectors of SD sampled from noise showing disentanglement. The dots and the images with borders show the resulting SD latent vectors and their corresponding "neutral" images with the text prompt - `Normal chest x-ray with no significant findings`. For each generated sample, we swap the original text condition to `Chest x-ray showing Support Devices` for one trajectory and to `Chest x-ray showing Pleural Effusion` for a different trajectory at multiple denoising steps during reverse diffusion.

respond to specific attributes. The direction along these trajectories can be discovered using an input text prompt. As we traverse in these directions, the desired attribute becomes more prominent in the image without affecting the other confounding attributes, indicating the disentanglement of stable diffusion latent space. This can be seen in Figure 4. To validate our hypothesis further, we also perform a curve fitting through these latent points and sample new points on this interpolated curve. The properties of the images from these interpolated points affirm disentangled trajectories for multiple attributes (shown in the Appendix).

Quantitative evaluations: In Table 2, the CFRT and the LPIPS scores between the "neutral image" and the conditionally generated style images at different timesteps along

Table 1. Performance of the disease or artifact classifiers

	CheXpert		ISIC				
	Support Devices	Pleural Effusion	MEL / NV	Hair	Gel Bubbles	Ink	Ruler
Accuracy	0.86	0.80	0.91	0.93	0.94	0.96	0.97
F1-score	0.88	0.79	0.88	0.91	0.78	0.89	0.88

Table 2. Quantitative results to evaluate the synthesized images (2500 samples per sub-class). Note: Neutral images for CheXpert are always healthy subjects without support devices. MEL and NV indicate the "neutral image" with MEL or NV.

	CheXpert		ISIC				
Style→	Pleural Effusion	Support Devices		Hair	Gel Bubbles	Ink	Ruler
CFRT↑	0.78	0.89	MEL	0.91	0.99	0.59	0.74
			NV	0.86	0.97	0.71	0.95
LPIPS↓	0.24	0.05	MEL	0.08	0.09	0.12	0.11
			NV	0.05	0.09	0.06	0.10
Interpolations							
CFRT↑	0.73	0.86	MEL	0.88	0.99	0.62	0.79
			NV	0.93	0.99	0.72	0.97
LPIPS↓	0.22	0.04	MEL	0.05	0.08	0.09	0.08
			NV	0.04	0.07	0.04	0.07

the trajectory (including interpolated samples) are high and low, respectively. This indicates that the attribute of interest along a trajectory changes more than other attributes, and all the interpolated samples show visual similarity to the neutral image. Table 1 shows the performance of the classifier Efficient-Net [31] for measuring CFRT score. Additionally, the cosine similarities, in Figure 5, indicate the non-linearity of trajectories for different attributes.

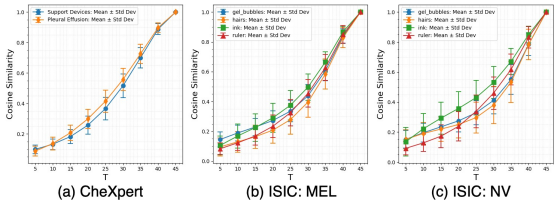

Figure 5. Cosine similarity for all the attributes.

4. Conclusion

Vision-language foundation models have rich latent representations that can be leveraged in medical imaging, where the data are limited. In this paper, we presented the first text-guided traversal along a non-linear trajectory in the disentangled latent space based on vision-language foundation models for medical images. The qualitative and quantitative results demonstrate that our method enables precise control over the synthesized images (including the interpolated images), allowing for targeted manipulation of visual attributes while preserving content. Future work will investigate ways to impose more structure and compositionality on the latent spaces.

Acknowledgements

The authors are grateful for funding provided by the Natural Sciences and Engineering Research Council of Canada, the Canadian Institute for Advanced Research (CIFAR) Artificial Intelligence Chairs program, Mila - Quebec AI Institute, Google Research, Calcul Québec, and the Digital Research Alliance of Canada.

References

[1] Alceu Bissoto, Catarina Barata, Eduardo Valle, and Sandra Avila. Artifact-based domain generalization of skin lesion models. In *European Conference on Computer Vision*, pages 133–149. Springer, 2022. 3

[2] Ricky TQ Chen, Xuechen Li, Roger B Grosse, and David K Duvenaud. Isolating sources of disentanglement in variational autoencoders. *Advances in neural information processing systems*, 31, 2018. 1

[3] Wei Chen and Mark Fuge. B\'eziergan: Automatic generation of smooth curves from interpretable low-dimensional parameters. *arXiv preprint arXiv:1808.08871*, 2018. 3

[4] Xi Chen, Yan Duan, Rein Houthooft, John Schulman, Ilya Sutskever, and Pieter Abbeel. Infogan: Interpretable representation learning by information maximizing generative adversarial nets. *Advances in neural information processing systems*, 29, 2016. 1

[5] Noel CF Codella, David Gutman, M Emre Celebi, Brian Helba, Michael A Marchetti, Stephen W Dusza, Aadi Kalloo, Konstantinos Liopyris, Nabin Mishra, Harald Kittler, et al. Skin lesion analysis toward melanoma detection: A challenge at the 2017 international symposium on biomedical imaging (isbi), hosted by the international skin imaging collaboration (isic). In *2018 IEEE 15th international symposium on biomedical imaging (ISBI 2018)*, pages 168–172. IEEE, 2018. 3

[6] Marc Combalia, Noel CF Codella, Veronica Rotemberg, Brian Helba, Veronica Vilaplana, Ofer Reiter, Cristina Carrera, Alicia Barreiro, Allan C Halpern, Susana Puig, et al. Bcn20000: Dermoscopic lesions in the wild. *arXiv preprint arXiv:1908.02288*, 2019. 3

[7] Erik Härkönen, Aaron Hertzmann, Jaakko Lehtinen, and Sylvain Paris. Ganspace: Discovering interpretable gan controls. *Advances in neural information processing systems*, 33:9841–9850, 2020. 2

[8] Irina Higgins, Loic Matthey, Arka Pal, Christopher P Burgess, Xavier Glorot, Matthew M Botvinick, Shakir Mohamed, and Alexander Lerchner. beta-vae: Learning basic visual concepts with a constrained variational framework. *ICLR (Poster)*, 3, 2017. 1

[9] Xun Huang and Serge Belongie. Arbitrary style transfer in real-time with adaptive instance normalization. In *Proceedings of the IEEE international conference on computer vision*, pages 1501–1510, 2017. 2

[10] Jeremy Irvin, Pranav Rajpurkar, Michael Ko, Yifan Yu, Silviana Ciurea-Ilcus, Chris Chute, Henrik Marklund, Behzad Haghgoo, Robyn Ball, Katie Shpanskaya, et al. Chexpert: A large chest radiograph dataset with uncertainty labels and expert comparison. In *Proceedings of the AAAI conference on artificial intelligence*, pages 590–597, 2019. 3

[11] Gwanghyun Kim, Taesung Kwon, and Jong Chul Ye. Diffusionclip: Text-guided diffusion models for robust image manipulation. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2426–2435, 2022. 3

[12] Hyunjik Kim and Andriy Mnih. Disentangling by factorising. In *International conference on machine learning*, pages 2649–2658. PMLR, 2018. 1

[13] Amar Kumar, Anita Kriz, Mohammad Havaei, and Tal Arbel. Prism: High-resolution & precise counterfactual medical image generation using language-guided stable diffusion. *MIDL*, 2025. 2

[14] Oran Lang, Yossi Gandelsman, Michal Yarom, Yoav Wald, Gal Elidan, Avinatan Hassidim, William T Freeman, Phillip Isola, Amir Globerson, Michal Irani, et al. Explaining in style: training a gan to explain a classifier in stylespace. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 693–702, 2021. 3

[15] F Leeb, Y Annadani, S Bauer, and B Schölkopf. Structured representation learning using structural autoencoders and hybridization. *arXiv preprint arXiv:2006.07796*, 2020. 1

[16] Luping Liu, Yi Ren, Zhijie Lin, and Zhou Zhao. Pseudo numerical methods for diffusion models on manifolds. *arXiv preprint arXiv:2202.09778*, 2022. 3

[17] Yang Liu, Eunice Jun, Qisheng Li, and Jeffrey Heer. Latent space cartography: Visual analysis of vector space embeddings. In *Computer graphics forum*, pages 67–78. Wiley Online Library, 2019. 3

[18] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael Jordan. Learning transferable features with deep adaptation networks. In *International conference on machine learning*, pages 97–105. PMLR, 2015. 1

[19] Qingjie Meng, Daniel Rueckert, and Bernhard Kainz. Learning cross-domain generalizable features by representation disentanglement. *arXiv preprint arXiv:2003.00321*, 2020. 1

[20] Michael E Mortenson. *Mathematics for computer graphics applications*. Industrial Press Inc., 1999. 3

[21] Taesung Park, Ming-Yu Liu, Ting-Chun Wang, and Jun-Yan Zhu. Semantic image synthesis with spatially-adaptive normalization. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 2337–2346, 2019. 2

[22] Nick Pawlowski, Daniel Coelho de Castro, and Ben Glocker. Deep structural causal models for tractable counterfactual inference. *Advances in neural information processing systems*, 33:857–869, 2020. 1

[23] Ethan Perez, Florian Strub, Harm De Vries, Vincent Dumoulin, and Aaron Courville. Film: Visual reasoning with a general conditioning layer. In *Proceedings of the AAAI conference on artificial intelligence*, 2018. 2

[24] Joaquin Quiñonero-Candela, Masashi Sugiyama, Anton Schwaighofer, and Neil D Lawrence. *Dataset shift in machine learning*. Mit Press, 2022. 1

[25] 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 (CVPR)*, pages 10684–10695, 2022. 2, 3

[26] Yujun Shen and Bolei Zhou. Closed-form factorization of latent semantics in gans. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1532–1540, 2021. 2

[27] Yujun Shen, Jinjin Gu, Xiaoou Tang, and Bolei Zhou. Interpreting the latent space of gans for semantic face editing. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9243–9252, 2020.

[28] Yujun Shen, Ceyuan Yang, Xiaoou Tang, and Bolei Zhou. Interfacegan: Interpreting the disentangled face representation learned by gans. *IEEE transactions on pattern analysis and machine intelligence*, 44(4):2004–2018, 2020. 2

[29] Christopher J Soelistyo and Alan R Lowe. Discovering interpretable models of scientific image data with deep learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 6884–6893, 2024. 1

[30] Jiaming Song, Chenlin Meng, and Stefano Ermon. Denoising diffusion implicit models. *arXiv preprint arXiv:2010.02502*, 2020. 2

[31] Mingxing Tan and Quoc Le. Efficientnet: Rethinking model scaling for convolutional neural networks. In *International conference on machine learning*, pages 6105–6114. PMLR, 2019. 4

[32] Youbao Tang, Yuxing Tang, Yingying Zhu, Jing Xiao, and Ronald M Summers. A disentangled generative model for disease decomposition in chest x-rays via normal image synthesis. *Medical Image Analysis*, 67:101839, 2021. 2

[33] Philipp Tschandl, Cliff Rosendahl, and Harald Kittler. The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. *Scientific data*, 5(1):1–9, 2018. 3

[34] Eric Tzeng, Judy Hoffman, Kate Saenko, and Trevor Darrell. Adversarial discriminative domain adaptation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 7167–7176, 2017. 1

[35] Qiucheng Wu, Yujian Liu, Handong Zhao, Ajinkya Kale, Trung Bui, Tong Yu, Zhe Lin, Yang Zhang, and Shiyu Chang. Uncovering the disentanglement capability in text-to-image diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 1900–1910, 2023. 2, 4

[36] Yankun Wu, Yuta Nakashima, and Noa Garcia. Not only generative art: Stable diffusion for content-style disentanglement in art analysis. In *Proceedings of the 2023 ACM International conference on multimedia retrieval*, pages 199–208, 2023. 2

[37] 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. 3

[38] Lei Zhou, Joseph Bae, Huidong Liu, Gagandeep Singh, Jeremy Green, Amit Gupta, Dimitris Samaras, and Prateek Prasanna. Lung swapping autoencoder: Learning a disentangled structure-texture representation of chest radiographs. *arXiv preprint arXiv:2201.07344*, 2022. 2