RL4Med-DDPO: Reinforcement Learning for Controlled Guidance Towards Diverse Medical Image Generation using Vision-Language Foundation Models

Parham Saremi*†[1,2], Amar Kumar†[1,2], Mohamed Mohamed[1,2], Zahra TehraniNasab[1,2], and Tal Arbel[1,2]

[1] Center for Intelligent Machines, McGill University, Montreal, Canada
[2] Mila - Quebec AI institute, Montreal, Canada
parham.saremi@mail.mcgill.ca
† equal contribution

Abstract. Vision-Language Foundation Models (VLFM) have shown a tremendous increase in performance in terms of generating high-resolution, photorealistic natural images. While VLFMs show a rich understanding of semantic content across modalities, they often struggle with fine-grained alignment tasks that require precise correspondence between image regions and textual descriptions, a limitation in medical imaging, where accurate localization and detection of clinical features are essential for diagnosis and analysis. To address this issue, we propose a multi-stage architecture where a pre-trained VLFM (e.g. Stable Diffusion) provides a cursory semantic understanding, while a reinforcement learning (RL) algorithm refines the alignment through an iterative process that optimizes for understanding semantic context. The reward signal is designed to align the semantic information of the text with synthesized images. Experiments on the public ISIC2019 skin lesion dataset demonstrate that the proposed method improves (a) the quality of the generated images, and (b) the alignment with the text prompt over the original fine-tuned Stable Diffusion baseline. We also show that the synthesized samples could be used to improve disease classifier performance for underrepresented subgroups through augmentation. Our code is accessible through the project website.[3]

Keywords: Medical Image Generation · Policy Optimization · Reinforcement Learning · Vision-Language Foundation Models.

1 Introduction

The development of state-of-the-art Vision-Language Foundation models (VLFM), such as Stable Diffusion [25], has significantly improved the image generation

* Corresponding author.
[3] https://parhamsaremi.github.io/rl4med-ddpo

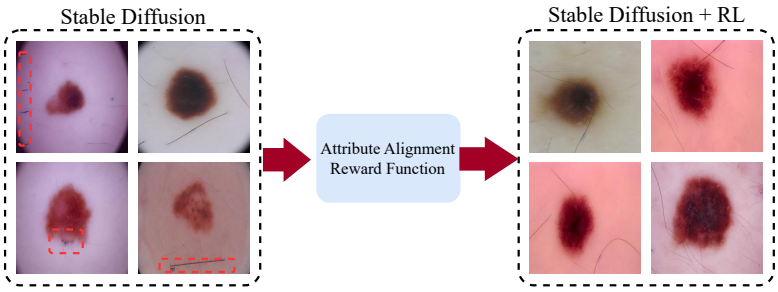

Fig. 1: Comparison of synthetic samples generated from Stable Diffusion (left image) and Stable Diffusion with Reinforcement Learning (right image). The text prompt for these image samples was - `A dermatoscopic image with melanoma showing hairs`. Note the unwanted but relevant artifacts that do not align with the input text prompt.

quality and resolution significantly over traditional generative models, such as VAEs and GANs [12,4]. In medical imaging, these new foundation models have demonstrated capability to generate highly realistic 2D images with fine details and textures. However, diffusion models inherit and amplify data bias [1,24] from large-scale training data, showing undesired behaviors. For example, when given a text prompt `A dermatoscopic image with melanoma showing hairs` to Stable Diffusion fine-tuned on skin cancer data, it generates realistic images with hairs. However, the synthesized images typically also contain well-known artifacts in the dataset, such as a ruler or ink shown in Figure 1. Thus, a semantic alignment mismatch exists between the text and the synthesized image.

Synthetic image generation is of importance, especially in medical imaging, as it can be used for tasks such as data augmentation [18], debiasing classifiers [19] or accurate detection and diagnosis of disease [13,33]. Additionally, high-resolution and precise image-generation capabilities in complex settings, such as drug discovery or personalized diagnosis, require analyzing counterfactual "what if" scenarios [11,20,23]. Recently, Denoising Diffusion models (DDMs) [15] have shown outstanding performances in high-resolution conditional image generation. Despite their impressive capability to synthesize images, they are prone to biases. These diffusion architectures utilize a controlled sampling process, which can either be classifier-free [16] or classifier-guided [8]. A promising alternative recently proposed for this goal is the use of Reinforcement Learning (RL) to optimize the diffusion process for improved control and adaptability [3,10,22,35]. Fine-tuning the diffusion model to optimize a desired reward function can enable these models to incorporate task-specific preferences, potentially reducing bias and improving alignment between generated samples and predefined constraints. Denoising Diffusion Policy Optimization (DDPO) [3] is an RL-based method that reframes the diffusion process as a multi-step Markov Decision Process (MDP) to optimize a given reward function. With the rise of policy-based methods, the effectiveness of various reward functions has been demonstrated in natural imag-

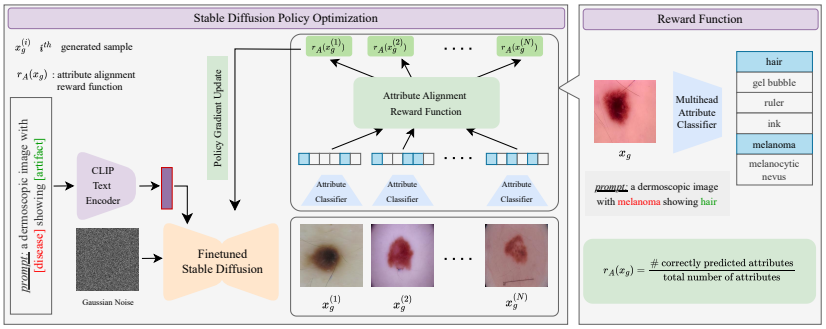

Fig. 2: Proposed architecture for policy optimization using a reward function for diverse and realistic image generation using fine-tuned Stable Diffusion. Given an input text prompt, the model synthesizes a realistic image x_g during the reverse diffusion. This generated image is then passed to a pre-trained classifier to compute the reward which helps guide the denoising UNet to improve image synthesis so it is better semantically aligned with the input text.

ing domains, including diversity-based rewards [22,35], alignment rewards [3], and visual rewards such as aesthetic quality [3]. However, their applicability in medical imaging remains underexplored.

In this work, we introduce the first framework for improved performance of text-guided image generation in medical imaging using Stable Diffusion through policy optimization in reinforcement learning. Specifically, we demonstrate that policy-based optimization improves the alignment between the input text prompts and the generated images. We propose a new metric - *Artifact Prevalence Rate (APR)* to compute the presence of the desired attributes in the synthesized image. Extensive experiments are performed on a publicly available dataset, ISIC2019 [5,6,31], demonstrating that the method is capable of synthesizing photorealistic images that are completely in alignment with the medical context. This permits more robust and bias-aware medical image synthesis in general, and specifically complements previous work [2,32] on ISIC which focus on mitigating biases for the task of image classification.

2 Methodology

The reinforcement learning framework combines Stable Diffusion in two stages: (i) Fine-tune the original Stable Diffusion v1.5 [25] with the medical dataset to align text-image pairs; (ii) Use a pre-trained classifier to compute the reward and update the weights of fine-tuned Stable Diffusion from stage (i), by performing policy optimization. The general framework of our method is now described and illustrated in Figure 2.

2.1 Denoising Diffusion Policy Optimization (DDPO)

Diffusion models: Markov chains [28] model the data generation process by gradually adding and removing noise. The forward process transforms input data x_0 into Gaussian noise x_T over T steps using a variance schedule β_t:

$$q(x_t|x_{t-1}) = \mathcal{N}(\sqrt{1-\beta_t}x_{t-1}, \beta_t I)$$

Defining $\alpha_t = 1 - \beta_t$ and $\bar{\alpha_t} = \prod_{s=1}^{t} \alpha_s$, we can directly express x_t as:

$$q(x_t|x_0) = \mathcal{N}(\sqrt{\bar{\alpha_t}}x_0, (1-\bar{\alpha_t})I)$$

The reverse diffusion process in DDPMs [15] removes noise using a trained model $p_\theta(x_{t-1}|x_t)$, approximating the true posterior $q(x_{t-1}|x_t, x_0)$. The model is trained via variational inference by minimizing the KL divergence between p and q.

DDPO: After the diffusion model is fine-tuned, it can be be further optimized to maximize the expected reward $J(\theta) = \mathbb{E}_{c\sim p(c), x_0\sim p_\theta(x_0|c)}[r(x_0, c)]$ where $p(c)$ is distribution over input text prompts/conditions and $p_\theta(x_0|c)$ is the sample distribution. This is achieved by first re-framing the diffusion model as a multi-step Markov Decision Process (MDP). An MDP can be defined by a set of states S, set of actions A, reward function R, and state transition distribution P. The RL agent tries to maximize the cumulative reward function by learning the policy $\pi(a_t|s_t)$, allowing it to select actions at each step.

Adapting the framework from [3], we define the MDP using the denoising model backward process as the policy:

$$\begin{aligned} s_t &\triangleq \{x_t, c, t\}, \quad a_t \triangleq x_{t-1} \\ \pi(a_t|s_t) &\triangleq p_\theta(x_{t-1}|x_t, c), \\ P(s_{t+1}|s_t, a_t) &\triangleq \{\delta(x_{t-1}), \delta(c), \delta(t-1)\} \\ \rho(s_0) &\triangleq \{\mathcal{N}(0,1), \delta(c), \delta(T)\} \\ R(s_t, a_t) &\triangleq \mathbb{1}\{t=0\} \cdot r(x_0, c), \end{aligned} \tag{1}$$

where the state s_t is defined as the combination of latent x_t, time t, and the condition c. Policy π for selecting the action is defined using the denoising model, and $\delta(x)$ denotes the Dirac function. While fine-tuning the diffusion model, we use the importance sampling estimator [17] which uses two models θ and θ' to calculate the policy gradients for θ:

$$\nabla_\theta L = \mathbb{E}\left[\sum_{t=0}^{T} \frac{p_\theta(x_{t-1}|x_t, c)}{p_{\theta'}(x_{t-1}|x_t, c)} \nabla_\theta \log p_\theta(x_{t-1}|x_t, c) r(x_0, c)\right]. \tag{2}$$

The expectation is taken over trajectories (intermediate latents in the denoising process) sampled from the previous model θ'. This estimator allows multiple gradient evaluations using samples generated from the old model θ'. However, as discussed in [3], the estimator's accuracy may degrade if p_θ and $p_{\theta'}$ diverge

Algorithm 1 SD Models with RL: implementing policy based gradient update.

Require: Pre-trained diffusion model $p_{\theta_{pre}}$, attribute classifier $f(\mathbf{X})$, reward function $r_A(c_{pred}, c)$.

1: Initialize $p_\theta = p_{\theta_{pre}}$
2: **while** θ not converged **do**
3: **for** each prompt $c \sim p(c)$ **do**
4: Sample M images $\mathbf{X}_g|c = \{x^1_{0,g}, \ldots, x^M_{0,g}\}, x^m_{0,g} \sim p_\theta(x_0|c)$, together with their intermediate states $x^m_{1:T,g}$.
5: Compute individual rewards with the attribute reward function for each condition c, $r_A(f(\mathbf{X}_g); c)$.
6: Take K rounds of policy gradient steps with 2.
7: **end for**
8: **end while**
9: **return** Fine-tuned diffusion model p_θ.

significantly. To address this issue, trust regions [26] are applied to limit the size of the update by using clipping techniques [27]. This will regularize the change of θ with respect to θ' resolving the problem [22].

2.2 Training Details

Finetuning Stable Diffusion: The original Stable Diffusion v1.5 [25] architecture consists of four main components: (1) image encoder, (2) Contrastive Language-Image Pre-training (CLIP) text encoder, (3) denoising U-Net for reverse diffusion and (4) an image decoder. Aside from the denoising U-Net, all other components of the Stable diffusion model remain frozen during fine-tuning. The unet is fine-tuned using the denoising score matching objective. Similar to PRISM [21], the input to the CLIP encoder is a template text: `A dermoscopic image with [disease] showing [artifact]`, where `disease` is melanoma or melanocytic nevis and `artifact` is hairs, gel bubbles, ink or ruler thus helping to synthesize semantically aligned images from the input text.
Policy based optimization: An H-head Efficient-Net [30] classifier ($H = 6$) is trained to identify the presence of `artifacts` including melanoma and melanocytic nevus. The reward function, $r_A(.)$, with respect to a generated image x_g is computed as a ratio of the number of attributes correctly predicted by the pre-trained Efficient-Net classifier to the total number of attributes. This reward function is then used to update the weights of denoising U-Net. This change in the value of the reward function between successive iterations is used as a stopping criterion for model fine-tuning. This is discussed in Algorithm 1.

2.3 Metrics & Evaluation of Synthesized Samples

The synthesized samples from our method, SD+RL, are compared against the fine-tuned Stable Diffusion (baseline), SD. Both these methods generate realistic samples, but the synthesized images aren't always perfectly aligned with the

Table 1: Summary of train, validation and test splits for ISIC 2019 dataset. Note that the prevalence of ink is significantly low compared to others

	Melanoma	Melanocytic Nevus	Hair	Gel Bubbles	Ink	Ruler
Train	2750	9254	4514	1300	201	1608
Validation	454	1665	802	228	34	308
Test	537	1956	989	257	37	341

input prompt. Thus, to measure the semantic alignment between image-text pairs, we propose a new metric, *Artifact Prevalence Rate (APR)*, computed as follows:

$$\text{APR} = \frac{\text{count of } x_i \in X : f(x_i) = C(\text{input text})}{N} \tag{3}$$

X are all the synthesized samples for the attribute under observation, N is the number of synthesized samples, $f(.)$ is the multi-head attribute classifier that identifies hair, gel bubbles, rule and ink, $C(.)$ is the function that one-hot encodes the `disease` and `artifact` information in the input text. A higher value of APR is expected as it would indicate that only the attribute mentioned in the text is prevalent in the synthesized samples and all others are ignored.

Finally, if the synthesized samples carry rich discriminative information about the domains, they should be able to improve the performance of subclasses with fewer real samples through augmentation. As such, the dataset was augmented with synthesized samples and the classifier's performance per subclass was evaluated before and after augmentation.

3 Experiments and Results

3.1 Dataset and Implementation Details

We perform experiments on a publicly available dataset, ISIC 2019 [5,6,31]. Table 1 shows the distribution of samples and the artifacts as per train, validation and test splits.

To evaluate and compare our method (RL+SD) with the baseline (SD), we use Fréchet Inception Distance (FID) [14] and LPIPS [34], in addition to the proposed APR metric. For a comprehensive analysis, we generate approximately 70K samples across various prompts and label combinations, ensuring robust evaluation of both models.

3.2 Qualitative Evaluations

Melanomas often have irregular, asymmetric shapes, while benign melanocytic nevus are typically well-circumscribed and symmetric [7,9]. Additionally, melanomas tend to have irregular, poorly defined borders and are often large with a diameter of about 6mm, while melanocytic nevus have clear, well-defined edges [29,9].

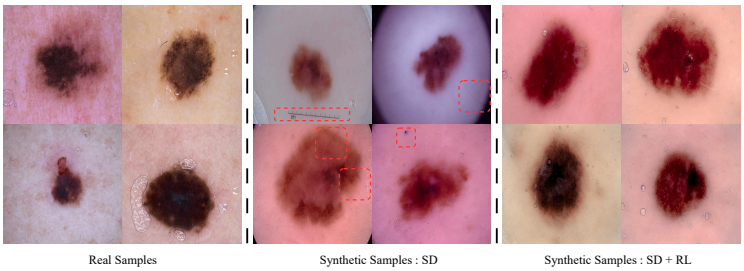

Fig. 3: Comparing real samples for the category "melanocytic nevus with gel bubbles" with the synthesized images using fine-tuned Stable Diffusion (SD) and fine-tuned Stable Diffusion with reinforcement learning (SD+RL). Note the unwanted artifacts present in the image synthesized by SD.

In Figure 3, both methods can accurately replicate the distinct melanomas and melanocytic nevus characteristics in the generated images. However, the samples from the method using SD+RL perform better as they create fewer or no unwanted attributes while synthesizing the images.

In Figure 4, we show that the proposed framework can create better samples for domains with no corresponding real samples i.e., during fine-tuning the Stable Diffusion model never saw these combinations of artifacts. For example, no real samples of melanocytic nevus with attributes hairs and ink are in the training data and SD+RL is able to synthesize visually better samples with better alignment and higher artifact quality.

3.3 Quantitative Evaluations

Table 2 evaluates the model using the APR metric across different disease and artifact combinations. The ability of our method (SD+RL) to maintain high

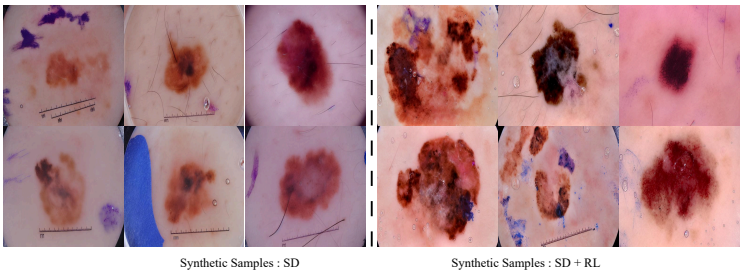

Fig. 4: Qualitative comparisons of synthesized images of subgroups (based on combinations of disease and artifacts) for which none or a few (less than 20) real samples are present. Note that some of these subgroups include combinations of attributes, such as melanoma with gel bubbles and ink or melanocytic nevus with ink and hair.

Table 2: Quantitative results for evaluating synthesized samples (1000 samples per class). Note the subgroups MEL with hair, ink and NV with ink have fewer real samples. Also, SD+RL consistently achieves higher APR values, even for classes with a limited number of real samples. (**Gel** = Gel Bubbles)

		Melanoma (MEL)				**Melanocytic Nevus (NV)**			
		Hair	**Gel**	**Ink**	**Ruler**	**Hair**	**Gel**	**Ink**	**Ruler**
# Real Sample		543	260	4	440	3051	489	21	520
APR ↑ (%)	SD (Baseline)	63.74	13.75	0.2	76.56	44.37	13.54	0.1	80
	SD+RL	**86.97**	**94.37**	**1.6**	**93.85**	**75.72**	**80.62**	**1.4**	**87.18**

APR values across underrepresented classes further validates the effectiveness of our optimization strategy. In Table 3, our method outperforms the pre-trained SD model with respect to APR, indicating that our method synthesizes images better aligned with the input prompts and generates fewer unwanted artifacts. Additionally, our methods also achieve a lower FID and similar LPIPS compared to the baseline. Finally, Table 4 presents the performance of the model after augmenting the training dataset with synthesized images. The results show that classifiers trained on real data augmented with our model's generated images achieve the highest performance. This confirms that our synthesized samples provide useful augmentations that enhance classifier accuracy, further validating the effectiveness of our approach in improving downstream tasks.

4 Conclusion

In this work, we demonstrate the first method showing alignment between the text prompt and image generation using a vision-foundation model guided by a policy optimization for medical imaging applications. We show through extensive qualitative and quantitative validation that these images align well with the input text prompt, and they are helpful for downstream tasks such as augmenting the classifier to improve performance over minority classes. Future work will explore the use of diverse policies for complex tasks such as subgroup clustering in the latent space for disease image marker discovery.

Table 3: Quantitative results for evaluating synthesized samples for 32 different prompts (1000 samples per prompt) of various artifacts with MEL or NV. LPIPS was calculated by randomly selecting 1000 samples per prompt and comparing them with corresponding real images, resulting in a total of 32,000 comparisons.

Model	**APR↑(%)**	**FID↓**	**LPIPS↓**
SD (Baseline)	18.28	121.47	0.60
SD+RL	**63.14**	**114.7**	**0.59**

Table 4: F1 and Accuracy for different attributes on ISIC test set for Real, RL-synthesized+Real, and SD-synthesized+Real. The metrics are calculated for each attribute independently of the others.

	Real		**SD-RL+Real**		SD+Real	
Setting	**F1**	**Accuracy**	**F1**	**Accuracy**	**F1**	**Accuracy**
Hair	0.91	93.10	**0.92**	**93.54**	0.91	93.26
Gel Bubbles	0.66	93.90	**0.70**	**94.06**	0.67	93.90
Ruler	0.89	96.95	**0.89**	**97.03**	0.85	96.23
Ink	**0.90**	**99.72**	0.89	99.68	0.86	99.60

Acknowledgments. 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 Quebec, and the Digital Research Alliance of Canada.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

References

1. Bansal, H., Yin, D., et al.: How well can text-to-image generative models understand ethical natural language interventions? arXiv preprint arXiv:2210.15230 (2022)
2. Bissoto, A., Valle, E., Avila, S.: Debiasing skin lesion datasets and models? not so fast. 2020 ieee. In: CVF Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 3192–3201 (2020)
3. Black, K., Janner, M., et al.: Training diffusion models with reinforcement learning. arXiv preprint arXiv:2305.13301 (2023)
4. Brock, A., Donahue, J., Simonyan, K.: Large scale gan training for high fidelity natural image synthesis. arXiv preprint arXiv:1809.11096 (2018)
5. Codella, N.C., Gutman, D., 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). pp. 168–172. IEEE (2018)
6. Combalia, M., Codella, N.C., et al.: Bcn20000: Dermoscopic lesions in the wild. arXiv preprint arXiv:1908.02288 (2019)
7. Damsky, W.E., Bosenberg, M.: Melanocytic nevi and melanoma: unraveling a complex relationship. Oncogene **36**(42), 5771–5792 (2017)
8. Dhariwal, P., Nichol, A.: Diffusion models beat gans on image synthesis. Advances in neural information processing systems **34**, 8780–8794 (2021)
9. Elder, D.E.: Precursors to melanoma and their mimics: nevi of special sites. Modern pathology **19**, S4–S20 (2006)
10. Fan, Y., Watkins, O., et al.: Dpok: Reinforcement learning for fine-tuning text-to-image diffusion models. Advances in Neural Information Processing Systems **36**, 79858–79885 (2023)

11. Favero, G.M., Saremi, P., et al.: Conditional diffusion models are medical image classifiers that provide explainability and uncertainty for free (2025), https://arxiv.org/abs/2502.03687
12. Gal, R., Patashnik, O., et al.: Stylegan-nada: Clip-guided domain adaptation of image generators. ACM Transactions on Graphics (TOG) **41**(4), 1–13 (2022)
13. Goceri, E.: Medical image data augmentation: techniques, comparisons and interpretations. Artificial Intelligence Review **56**(11), 12561–12605 (2023)
14. Heusel, M., Ramsauer, H., et al.: Gans trained by a two time-scale update rule converge to a local nash equilibrium. Advances in neural information processing systems **30** (2017)
15. Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. In: Advances in Neural Information Processing Systems. vol. 33, pp. 6840–6851 (2020)
16. Ho, J., Salimans, T.: Classifier-free diffusion guidance. arXiv preprint arXiv:2207.12598 (2022)
17. Kakade, S., Langford, J.: Approximately optimal approximate reinforcement learning. In: Proceedings of the nineteenth international conference on machine learning. pp. 267–274 (2002)
18. Kebaili, A., Lapuyade-Lahorgue, J., Ruan, S.: Deep learning approaches for data augmentation in medical imaging: a review. Journal of imaging **9**(4), 81 (2023)
19. Kumar, A., Fathi, N., et al.: Debiasing counterfactuals in the presence of spurious correlations. In: Workshop on Clinical Image-Based Procedures. pp. 276–286. Springer (2023)
20. Kumar, A., Hu, A., et al.: Counterfactual image synthesis for discovery of personalized predictive image markers. In: MICCAI Workshop on Medical Image Assisted Blomarkers' Discovery. pp. 113–124. Springer (2022)
21. Kumar, A., Kriz, A., et al.: Prism: High-resolution & precise counterfactual medical image generation using language-guided stable diffusion. MIDL (2025)
22. Miao, Z., Wang, J., et al.: Training diffusion models towards diverse image generation with reinforcement learning. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 10844–10853 (2024)
23. Pawlowski, N., Coelho de Castro, D., Glocker, B.: Deep structural causal models for tractable counterfactual inference. Advances in neural information processing systems **33**, 857–869 (2020)
24. Perera, M.V., Patel, V.M.: Analyzing bias in diffusion-based face generation models. In: 2023 IEEE International Joint Conference on Biometrics (IJCB). pp. 1–10. IEEE (2023)
25. Rombach, R., Blattmann, A., et al.: High-resolution image synthesis with latent diffusion models. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 10684–10695 (June 2022)
26. Schulman, J., Levine, S., et al.: Trust region policy optimization. In: International conference on machine learning. pp. 1889–1897. PMLR (2015)
27. Schulman, J., Wolski, F., et al.: Proximal policy optimization algorithms. arXiv preprint arXiv:1707.06347 (2017)
28. Sohl-Dickstein, J., Weiss, E.A., et al.: Deep unsupervised learning using nonequilibrium thermodynamics. Journal of Statistical Mechanics: Theory and Experiment **2015**(7), P07019 (2015)
29. Sung, W.W., Chang, C.H.: Nevi, dysplastic nevi, and melanoma: Molecular and immune mechanisms involving the progression. Tzu Chi Medical Journal **34**(1), 1–7 (2022)

30. Tan, M., Le, Q.: Efficientnet: Rethinking model scaling for convolutional neural networks. In: International conference on machine learning. pp. 6105–6114. PMLR (2019)
31. Tschandl, P., Rosendahl, C., Kittler, H.: The ham10000 dataset, a large collection of multi-source dermatoscopic images of common pigmented skin lesions. Scientific data **5**(1), 1–9 (2018)
32. Yan, S., Yu, Z., et al.: Towards trustable skin cancer diagnosis via rewriting model's decision. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition. pp. 11568–11577 (2023)
33. Zhang, C., Tavanapong, W., et al.: Real data augmentation for medical image classification. In: 6th Joint International Workshops, CVII-STENT 2017 and 2nd International Workshop, LABELS 2017, Held in Conjunction with MICCAI 2017, Québec City, QC, Canada, September 10–14, 2017, Proceedings 2. pp. 67–76. Springer (2017)
34. Zhang, R., Isola, P., et al.: The unreasonable effectiveness of deep features as a perceptual metric. In: Proceedings of the IEEE conference on computer vision and pattern recognition. pp. 586–595 (2018)
35. Zhang, Y., Tzeng, E., et al.: Large-scale reinforcement learning for diffusion models. In: European Conference on Computer Vision. pp. 1–17. Springer (2024)