VL-RewardGen: Vision–Language Reward Driven Skin Lesion Image Generation

Yusung Chu and Sejung Yang

Department of Precision Medicine, Wonju College of Medicine, Yonsei University, Wonju, Republic of Korea
dbtjd9968@yonsei.ac.kr, syang@yonsei.ac.kr

Abstract. Recent advances in reinforcement learning (RL)-based fine-tuning of diffusion models have improved controllability in image generation. However, these approaches typically depend on predefined reward functions or attribute-specific annotations, which remain scarce in many medical domains. In dermatology, fine-grained labels—such as dermoscopic attributes—are often limited, restricting the synthesis of clinically targeted image variations. To address this challenge, we propose a vision–language (VL) reward-driven framework for controllable skin lesion image generation, termed VL-RewardGen. Our method leverages MONET, a dermatology-specific pretrained VL model, to compute semantic alignment scores between generated images and concept texts describing lesion classes and visual attributes. These alignment scores serve as reward signals to fine-tune a diffusion model via RL, directly optimizing image–concept consistency without requiring additional fine-grained supervision. We evaluate VL-RewardGen on the ISIC 2019 and MILK-10K datasets. Compared to standard fine-tuned Stable Diffusion and a supervised-reward RL baseline, our method achieves improved controllability while maintaining high visual fidelity. Overall, VL-RewardGen enables semantically controlled dermatology image generation and demonstrates the potential of domain-specific VL models as scalable reward functions for targeted medical data synthesis. Our code is publicly available at https://github.com/wormschu/VL-RewardGen-Vision-Language-Reward-Driven-Skin-Lesion-Image-Generation.

Keywords: Skin lesion image · Medical Image Generation · Vision–Language Model · Reinforcement Learning.

1 Introduction

Generative models have rapidly matured into a practical tool for medical imaging, where limited data availability and expensive expert annotation frequently hinder the development of robust clinical AI systems [3, 11]. By synthesizing additional training samples, generative modeling can support data augmentation, reduce domain gaps, and help represent rare phenotypes that are under-sampled in real-world cohorts. Recent advances in diffusion models have further improved

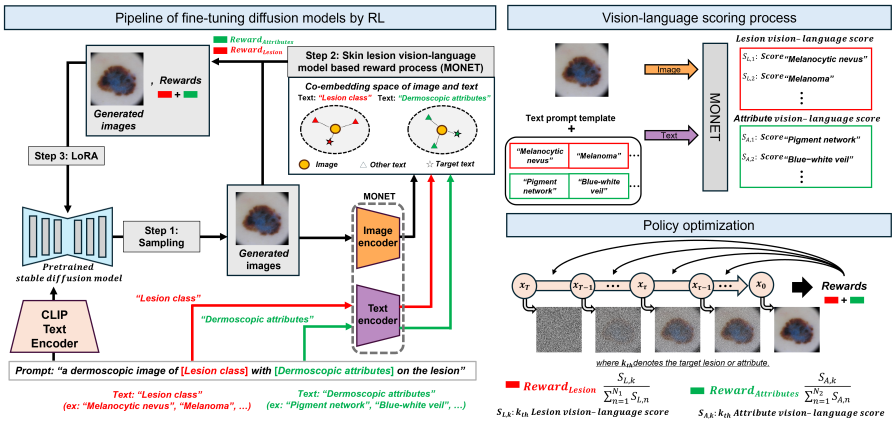

Fig. 1. Overview of VL-RewardGen. Given a generation prompt, Stable Diffusion samples images, and MONET computes concept-level vision–language alignment scores for the target lesion class and dermoscopic attributes. These scores are combined into lesion and attribute rewards and used to fine-tune the diffusion model via DDPO [1] with parameter-efficient LoRA updates [8].

image fidelity and diversity, enabling high-resolution synthesis and making diffusion a compelling backbone for medical image generation and augmentation [17, 12, 10].

However, clinically useful generation requires fine-grained control over subtle clinical attributes and their compositions [5]. In practice, fine-grained attribute supervision and sufficiently diverse attribute compositions are often limited, making prompt-faithful conditional generation challenging [2, 22].

To better align generated images with clinically meaningful targets, recent work has explored feedback-driven fine-tuning of diffusion models. Reward-based approaches such as DDPO [1] treat the denoising process as a policy and optimize sample-level objectives, but in medical imaging reliable reward specification typically requires clinically grounded criteria and expert involvement [21]. RL4Med-DDPO [18] instantiates this paradigm for dermoscopy by training task-specific supervised reward predictors, which can limit scalability to new attributes and acquisition settings.

To overcome this limitation, we present VL-RewardGen, a vision–language (VL) reward-driven reinforcement learning (RL) framework for controllable diffusion-based synthesis. Instead of relying on task-specific supervised reward classifiers (and the corresponding attribute-labeled train data), we compute rewards directly from a medical VL model that scores image–concept consistency in a shared embedding space, and optimize the diffusion generator via policy optimization.

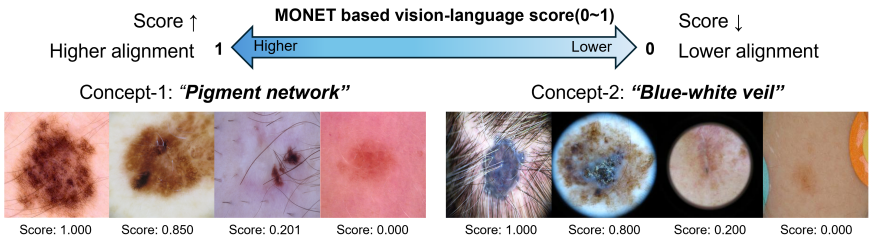

Fig. 2. Qualitative examples of MONET-based vision–language scores for dermoscopic attributes. For each concept, real dermoscopic images are ordered from high to low score (left to right). MONET outputs a normalized concept score $S(x, c) \in [0, 1]$ (Eq. 1), where higher values indicate stronger image–concept (i.e., *lesion class* or *dermoscopic attribute*) consistency.

Given a text prompt specifying target clinical concepts, the generated image and textual descriptors are encoded by the VL model, and their alignment score is used as the reward signal. We optimize the diffusion model using DDPO with parameter-efficient LoRA [8] updates, enabling scalable fine-grained control without training additional labeled reward predictors.

2 Methodology

As shown in Fig. 1, VL-RewardGen fine-tunes a diffusion generator using VL rewards. We initialize our generator from the dermoscopy-adapted Stable Diffusion (SD) [17] checkpoint used in RL4Med-DDPO [18] (i.e., without applying their subsequent reward-based policy optimization), and then adapt it via our VL model based reward policy optimization. Instead of computing rewards with a supervised attribute classifier, we compute rewards from a dermatology-specific VL model by scoring image–concept consistency in a shared embedding space.

2.1 Dermatology-specific vision–language model (MONET)

We use MONET [13], a dermatology-specific VL model, as a concept scoring function for reward computation. Given an image x and a concept text c (i.e., *lesion class* or *dermoscopic attribute*), MONET outputs a normalized concept score $S(x, c) \in [0, 1]$, where higher values indicate stronger evidence for concept c in image x. In this work, we use $S(x, c)$ as our VL alignment score.

Following the MONET scoring protocol, we compute $S(x, c)$ by averaging template-wise softmax scores between a target prompt and a template-specific reference text:

$$S(x,c) = \frac{1}{K}\sum_{k=1}^{K} \frac{\exp(\langle t(\tau_k(c)), v(x)\rangle/\gamma)}{\exp(\langle t(\tau_k(c)), v(x)\rangle/\gamma) + \exp(\langle t(\rho_k), v(x)\rangle/\gamma)}, \tag{1}$$

where $v(x)$ and $t(\cdot)$ are ℓ_2-normalized image and text embeddings, $\{\tau_k\}_{k=1}^K$ are prompt templates, $\{\rho_k\}_{k=1}^K$ are corresponding reference texts, and γ is a temperature.

Fig. 2 provides qualitative examples of concept scores for dermoscopic attributes. Images that exhibit the target attribute (e.g., *pigment network* or *blue-white veil*) obtain higher scores, whereas images without the attribute obtain lower scores. In VL-RewardGen, we use these concept scores as rewards for diffusion model policy optimization.

2.2 Vision–language reward-based policy optimization

Reward design: We compute rewards from MONET concept scores for dermoscopic attributes and lesion type. Since lesion types (NV vs. MEL) are mutually exclusive for a given image, we use a competition-based lesion reward. In contrast, dermoscopic attributes may be absent or co-occur, and we define the attribute reward to handle none/single/multiple targets.

Attribute reward: Given a prompt, we extract a target attribute set $A^+ \subseteq \mathcal{A}_{\text{derm}}$, where $\mathcal{A}_{\text{derm}}$ denotes the dermoscopic attribute pool and $S(x,a) \in [0,1]$ is the MONET score (Eq. 1). We define the reward to (i) suppress attributes when none are desired, (ii) promote a single target while discouraging spurious co-occurrence, and (iii) encourage simultaneous expression when multiple attributes are requested:

$$r_{\text{attr}}(x; A^+) = \begin{cases} \frac{1}{|\mathcal{A}_{\text{derm}}|} \sum\limits_{a \in \mathcal{A}_{\text{derm}}} \mathbb{1}[S(x,a) \le \tau], & A^+ = \varnothing \quad \text{(i)} \\ \frac{S(x,a)}{\sum\limits_{a' \in \mathcal{A}_{\text{derm}}} S(x,a') + \varepsilon}, & A^+ = \{a\} \quad \text{(ii)} \\ \frac{1}{|A^+|} \sum\limits_{a \in A^+} S(x,a), & |A^+| > 1 \quad \text{(iii)} \end{cases} \tag{2}$$

where ε is a small constant and τ is a threshold for the *none-attribute* setting.

Lesion reward and final reward: In addition to dermoscopic attributes, we enforce lesion-type consistency using MONET scores for the lesion classes $\mathcal{C} = \{\text{melanocytic nevus (NV)}, \text{melanoma (MEL)}\}$. Given the target lesion class $\ell \in \mathcal{C}$, we define a ratio-style lesion reward:

$$r_{\text{lesion}}(x; \ell) = \frac{S(x,\ell)}{\sum\limits_{c \in \mathcal{C}} S(x,c) + \varepsilon}, \tag{3}$$

where ε is a small constant.

The final reward is computed as a weighted sum of lesion and attribute rewards:

$$r(x; \ell, A^+) = \gamma\, r_{\text{lesion}}(x; \ell) + (1-\gamma)\, r_{\text{attr}}(x; A^+). \tag{4}$$

DDPO optimization: We fine-tune the diffusion model using DDPO, which formulates the denoising trajectory as a policy and maximizes the expected reward over generated samples [1]. We use a PPO-style clipped policy-gradient objective for stable updates. During optimization, we update only the LoRA parameters in the denoising U-Net while keeping the VAE and text encoder frozen.

2.3 Training details

We enable LoRA [8] and optimize only the LoRA parameters in the denoising U-Net, while keeping the VAE and text encoder frozen. Training is performed for 50 epochs using mixed precision (FP16). Each epoch samples 8 batches of 8 images (64 samples) with DDIM [19] sampling (50 steps, guidance scale 5.0, $\eta = 1.0$), followed by policy optimization with a per-GPU training batch size of 2 and gradient accumulation of 16 steps. We use AdamW with learning rate 5×10^{-5} and clip gradients at 1.0. Advantages are normalized using per-prompt statistic tracking (buffer size 32, minimum count 16). We set the none-attribute threshold to $\tau = 0.25$ and the final reward weighting to $\gamma = 0.5$ (Eq. 4). All experiments were conducted on a single NVIDIA RTX A6000 GPU.

3 Experiments and Results

3.1 Experimental setup

We evaluate controllable generation for dermoscopic attributes using two representative concepts: *pigment network* and *blue-white veil*. Prompts follow:

`a dermoscopic image of {lesion type} with {attributes} on the lesion`,

where {lesion type} $\in$ {*NV*, *MEL*} and {attributes} is one of: (i) none, (ii) a single attribute, or (iii) both attributes simultaneously. This yields eight prompt conditions (2 lesion types $\times$ 4 attribute settings).

To enable a fair comparison to classifier-based RL fine-tuning, we train the RL4Med-DDPO reward classifier using attribute supervision from Derm7pt [9] and report RL4Med-DDPO as a baseline under the same prompt conditions.

3.2 Datasets

We evaluate VL-RewardGen on two skin-lesion open datasets, ISIC 2019 [20, 4, 7] and MILK-10K [15], focusing on the NV and MEL categories. For evaluation, we use the ISIC 2019 Challenge test split with 3,822 images (NV: 2,495, MEL: 1,327) and a MILK-10K subset with 1,196 images (NV: 746, MEL: 450).

Since fine-grained dermoscopic attributes are not exhaustively annotated in these datasets, we construct attribute-specific evaluation subsets using MONET-based filtering. For each target attribute, we select real images whose MONET concept score satisfies $S(x, a) \geq 0.7$ and use the resulting subset as the reference

Table 1. Dermoscopic attribute controllability. FID/KID (↓) on ISIC 2019 and MILK-10K. Attr-1 denotes *blue-white veil* and Attr-2 denotes *pigment network*. Best and second-best results per column are highlighted in **bold** and underline, respectively.

	No attr FID/KID↓	Attr-1 FID/KID↓	Attr-2 FID/KID↓	Attr-1+Attr-2 FID/KID↓
ISIC 2019 dataset				
SD	18.77/9.31	26.34/14.44	16.64/9.77	34.83/18.00
+RL4Med-DDPO [18]	18.73/9.65	25.24/14.71	16.81/10.09	33.74/17.82
+VL-RewardGen (Cutoff$_{0.5}$)	17.07/7.20	23.17/11.30	17.09/9.56	30.96/**17.05**
+VL-RewardGen (proposed)	**16.40**/**7.18**	**22.25**/**11.26**	**15.82**/**9.53**	**30.88**/17.06
MILK-10K dataset				
SD	22.70/14.91	20.56/18.28	19.84/15.59	21.26/18.57
+RL4Med-DDPO [18]	23.65/15.83	20.86/18.66	19.47/15.57	21.50/**18.40**
+VL-RewardGen (Cutoff$_{0.5}$)	**22.00**/14.14	20.58/18.64	20.35/**15.31**	21.25/19.89
+VL-RewardGen (proposed)	22.77/**13.35**	**20.30**/**17.54**	**19.45**/15.53	**21.21**/18.48

distribution for quality evaluation. For each prompt, we generate 200 synthetic samples (1,600 samples in total across the eight prompt conditions). We report Fréchet Inception Distance (FID) and Kernel Inception Distance (KID) between generated samples and the corresponding real subsets to assess generation quality and realism.

3.3 Quantitative Results

In Table 1, we report FID and KID (↓) for dermoscopic attribute controllability on ISIC 2019 and MILK-10K under four prompt settings. Overall, VL-RewardGen achieves lower FID/KID than the fine-tuned Stable Diffusion (SD) baseline across most settings, indicating that our method synthesizes samples that are closer to the real data distribution while better satisfying attribute-level constraints. Compared to RL4Med-DDPO with a supervised reward classifier, our approach shows improved or comparable FID/KID in the majority of conditions, with particularly strong gains in the co-occurrence setting, highlighting the benefit of vision–language rewards for multi-attribute control.

We additionally include VL-RewardGen (Cutoff$_{0.5}$), which uses a commonly adopted thresholded reward by treating MONET scores as multi-label predictions ($S(x, a) > 0.5$) to form a classifier-style reward. While it improves over SD, the proposed reward yields better or comparable FID/KID, indicating stronger guidance than cutoff-based supervision.

3.4 Visualization Results

Fig. 3 qualitatively compares attribute-controlled generation across baselines. VL-RewardGen produces samples with clearer and more consistent expression of the requested dermoscopic patterns than the SD baseline and RL4Med-DDPO, particularly in the co-occurrence setting. We also observe fewer spurious patterns, suggesting improved prompt-faithful controllability under fine-grained attribute constraints.

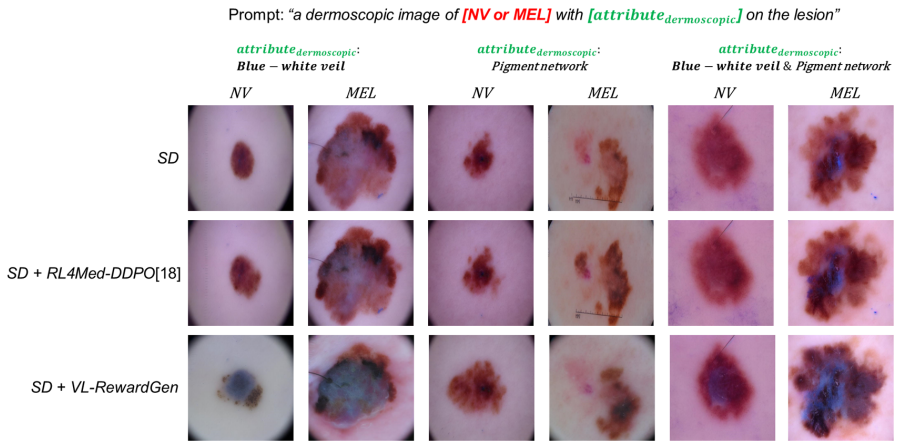

Fig. 3. Qualitative dermoscopic attribute controllability. Samples generated under prompts of the form. Columns correspond to target attribute settings (blue-white veil, pigment network, and their co-occurrence), and rows compare Fine-tuned SD, RL4Med-DDPO (supervised reward classifier), and VL-RewardGen.

3.5 Downstream Classification with Data Augmentation

To assess the utility of synthesized images for downstream learning, we perform a MEL vs. NV classification experiment using synthetic augmentation. We first construct a balanced real training set with 600 MEL and 600 NV images (1,200 total), randomly sampled from the ISIC 2019 training split. For each attribute setting, we augment this set by incorporating 200 synthetic MEL and 200 synthetic NV samples (400 total) generated by each method. We then fine-tune a ViT-B/16 classifier [6] on the resulting training set and evaluate on the ISIC 2019 Challenge test split.

Fig. 4 presents accuracy and F1-score. The dashed line denotes the real-data baseline, and stars indicate the strongest augmentation method for each attribute setting. Overall, augmentation with VL-RewardGen consistently improves performance over the real-only baseline and yields the strongest or comparable gains across attribute conditions.

3.6 Ablation Study: Vision–language reward model

In Table 2, we ablate the vision–language model used for reward computation by replacing MONET with CLIP [16], a general-purpose vision–language model pretrained on large-scale web data, and PMC-CLIP [14], a biomedical-domain CLIP-style model pretrained on PubMed Central image–caption pairs. Overall, MONET performs best in most settings across ISIC 2019 and MILK-10K, supporting the use of a dermatology-specific VL model as a more effective reward signal for dermoscopic attribute control.

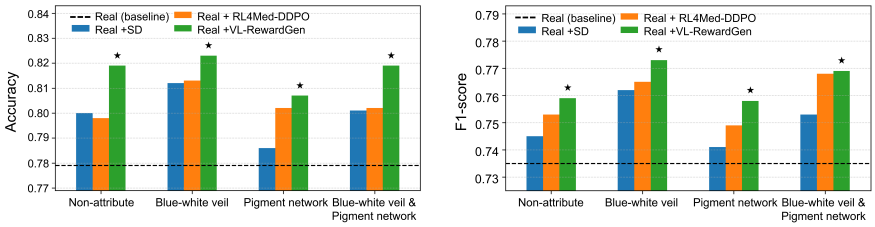

Fig. 4. Downstream classification gains from augmentation. The dashed line indicates the real-data baseline, and stars denote the best augmentation method for each attribute setting.

Table 2. VL reward model ablation FID/KID (↓) when replacing MONET with CLIP and PMC-CLIP on ISIC 2019 and MILK-10K. Attr-1 denotes *blue-white veil* and Attr-2 denotes *pigment network*. Best and second-best results per column are highlighted in **bold** and underline, respectively.

	No attr FID/KID↓	Attr-1 FID/KID↓	Attr-2 FID/KID↓	Attr-1+Attr-2 FID/KID↓
ISIC 2019 dataset				
CLIP [16]	18.62/8.93	26.97/14.67	17.36/9.70	34.61/17.79
PMC-CLIP [14]	18.23/9.29	27.70/15.38	17.41/9.80	34.39/17.58
MONET [13]	**16.40**/**7.18**	**22.25**/**11.26**	**15.82**/**9.53**	**30.88**/**17.06**
MILK-10K dataset				
CLIP [16]	23.35/14.47	22.67/19.05	19.64/15.71	22.43/18.78
PMC-CLIP [14]	23.80/14.98	22.73/18.89	20.08/15.54	22.07/**17.80**
MONET [13]	**22.77**/**13.35**	**20.30**/**17.54**	**19.45**/**15.53**	**21.21**/18.48

4 Conclusion

We propose VL-RewardGen, a VL reward-driven framework for controllable dermoscopic image generation. By using MONET-derived concept scores as rewards within DDPO, our method improves fine-grained attribute controllability while maintaining generation fidelity on both ISIC 2019 and MILK-10K. Importantly, VL-RewardGen mitigates the reliance on task-specific supervised reward predictors and additional attribute-labeled training data by leveraging a domain-specific VL model as a scalable reward function.

Limitations: Our approach depends on the calibration and concept coverage of the underlying VL reward model, and we currently validate controllability on a limited set of dermoscopic attributes and lesion categories. Because MONET-derived scores are also used to construct attribute-specific reference subsets, future studies should include independent expert annotations to reduce reward–evaluation coupling. Generated images are not intended to replace real patient data, but to serve as targeted augmentation for underrepresented concept combinations. We did not explicitly evaluate subgroup-level behavior across skin

color or broader clinical phenotypes. Future work will extend VL-RewardGen to richer attribute taxonomies, skin-tone-aware descriptors, and more rigorous expert-based and statistical validation.

Acknowledgments. This work was supported by the National Research Foundation of Korea (NRF) grants funded by the Korea government (MSIT) (Nos. RS-2025-25462906 and NRF-2022R1A2C2091160), the Bio&Medical Technology Development Program of the NRF funded by the Korea government (MSIT) (No. RS-2024-00440802), and the Basic Science Research Program through the NRF funded by the Ministry of Education (No. RS-2025-25433289).

Disclosure of Interests. The authors declare that they have no competing interests.

References

1. Black, K., Janner, M., Du, Y., Kostrikov, I., Levine, S.: Training diffusion models with reinforcement learning. arXiv preprint arXiv:2305.13301 (2023)
2. Chaichuk, M., Gautam, S., Hicks, S., Tutubalina, E.: Prompt to polyp: Clinically-aware medical image synthesis with diffusion models. arXiv preprint arXiv:2505.05573 (2025)
3. Chen, Y., Yang, X.H., Wei, Z., Heidari, A.A., Zheng, N., Li, Z., Chen, H., Hu, H., Zhou, Q., Guan, Q.: Generative adversarial networks in medical image augmentation: a review. Computers in Biology and Medicine **144**, 105382 (2022)
4. Codella, N.C., Gutman, D., Celebi, M.E., Helba, B., Marchetti, M.A., Dusza, S.W., Kalloo, A., Liopyris, K., Mishra, N., Kittler, H., 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)
5. Cong, Y., Min, M.R., Li, L.E., Rosenhahn, B., Yang, M.Y.: Attribute-centric compositional text-to-image generation. International Journal of Computer Vision **133**(7), 4555–4570 (2025)
6. Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al.: An image is worth 16x16 words: Transformers for image recognition at scale. arXiv preprint arXiv:2010.11929 (2020)
7. Hernández-Pérez, C., Combalia, M., Podlipnik, S., Codella, N.C., Rotemberg, V., Halpern, A.C., Reiter, O., Carrera, C., Barreiro, A., Helba, B., et al.: Bcn20000: Dermoscopic lesions in the wild. Scientific data **11**(1), 641 (2024)
8. Hu, E.J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W.: Lora: Low-rank adaptation of large language models. In: International Conference on Learning Representations (ICLR) (2022)
9. Kawahara, J., Daneshvar, S., Argenziano, G., Hamarneh, G.: Seven-point checklist and skin lesion classification using multitask multimodal neural nets. IEEE Journal of Biomedical and Health Informatics (2018)
10. Kazerouni, A., Aghdam, E.K., Heidari, M., Azad, R., Fayyaz, M., Hacihaliloglu, I., Merhof, D.: Diffusion models in medical imaging: A comprehensive survey. Medical image analysis **88**, 102846 (2023)

11. Khosravi, B., Purkayastha, S., Erickson, B.J., Trivedi, H.M., Gichoya, J.W.: Exploring the potential of generative artificial intelligence in medical image synthesis: opportunities, challenges, and future directions. The Lancet Digital Health (2025)
12. Kim, B., Ye, J.C.: Diffusion deformable model for 4d temporal medical image generation. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 539–548. Springer (2022)
13. Kim, C., Gadgil, S.U., DeGrave, A.J., et al.: Transparent medical image ai via an image–text foundation model grounded in medical literature. Nature Medicine (2024). https://doi.org/10.1038/s41591-024-02887-x
14. Lin, W., Zhao, Z., Zhang, X., Wu, C., Zhang, Y., Wang, Y., Xie, W.: Pmc-clip: Contrastive language-image pre-training using biomedical documents. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 525–536. Springer (2023)
15. Philipp, T., Bengü, A.N., Cliff, R., Veronica, R., Verche, T., Jochen, W., Katharina, W.A., Christoph, M., Nicholas, K., Allan, H., et al.: Milk10k: A hierarchical multimodal imaging learning toolkit for diagnosing pigmented and non-pigmented skin cancer and its simulators. Journal of Investigative Dermatology (2025)
16. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., et al.: Learning transferable visual models from natural language supervision. In: International conference on machine learning. pp. 8748–8763. PmLR (2021)
17. Rombach, R., Blattmann, A., Lorenz, D., Esser, P., Ommer, B.: High-resolution image synthesis with latent diffusion models. In: CVPR (2022)
18. Saremi, P., Kumar, A., Mohamed, M., TehraniNasab, Z., Arbel, T.: Rl4med-ddpo: reinforcement learning for controlled guidance towards diverse medical image generation using vision-language foundation models. In: International Conference on Medical Image Computing and Computer-Assisted Intervention. pp. 478–488. Springer (2025)
19. Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. In: International Conference on Learning Representations (ICLR) (2021), https://openreview.net/forum?id=St1giarCHLP
20. 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), 180161 (2018)
21. Wang, J., Zhang, Y., Ding, Z., Hamm, J.: Doctor approved: Generating medically accurate skin disease images through ai-expert feedback. arXiv preprint arXiv:2506.12323 (2025)
22. de Wilde, B., Saha, A., de Rooij, M., Huisman, H., Litjens, G.: Medical diffusion on a budget: Textual inversion for medical image generation. In: MIDL (2024)