Supplementary Material: Towards Facilitated Fairness Assessment of AI-based Skin Lesion Classifiers Through GenAI-based Image Synthesis

Ko Watanabe* Stanislav Frolov* Aya Hassan David Dembinsky

Adriano Lucieri Andreas Dengel

German Research Center for Artificial Intelligence (DFKI)

first.last@dfki.de

*Equal Contribution

Analysis of sampling parameters

To better understand the performance of our generative model we sweep the classifier-free guidance (CFG) scale and number of sampling steps, see Table 1. We find that the default sampling parameters are not optimal. Instead, a lower CFG and higher number of steps produce the best FID (lower is better). Therefore, we show synthetic images using these optimal parameters in the main paper. For comparison, we display synthetic images with default sampling parameters in Figure 1.

Impact of data availability on generation quality

Next, we are interested in the relationship between data availability and final generation quality. To that end, we measure the FID of synthetic images grouped by age against real data and compute the corresponding available training data for said age group in Table 2. We find that the generation quality correlates well with the data availability. Specifically, the best FID is achieved for age group 50 and 60, both represented more than 10% in the training data.

Visualization of real testing data

In Figure 2, we visualize real testing data from the MILK10k [2] dataset. Firstly, we observe that many combinations are simply not available thus making fairness assessment challenging. Secondly, as seen in both the real and synthetic images, they do not become progressively darker with higher Fitzpatrick skin types. Prior work [1, 3] shows that this intuitive assumption fails because lighting, camera settings, and lesion visibility heavily affect perceived darkness, leading expert-, crowd-, and algorithm-based annotations to diverge. Due to this variability, raw pixel intensity is an unreliable indicator of skin type, and higher categories do not consistently appear darker, highlighting the need for calibrated annotations rather than appearance alone.

Nearest neighbours visualization

To further assess the capabilities of our generative model, we visualize synthetic images alongside their real nearest neighbours from the training set in Figure 3–Figure 6. The generated samples are highly realistic yet clearly distinct from their nearest neighbours, indicating that the model does not merely memorize the training data.

Zero-shot classification comparison

Finally, we are interested in the performance of WhyLesionCLIP [4], a fine-tuned version of OpenCLIP, on real as well as our synthetic images, see Figure 7 to Figure 13. WhyLesionCLIP [4] captures dermatological and clinical

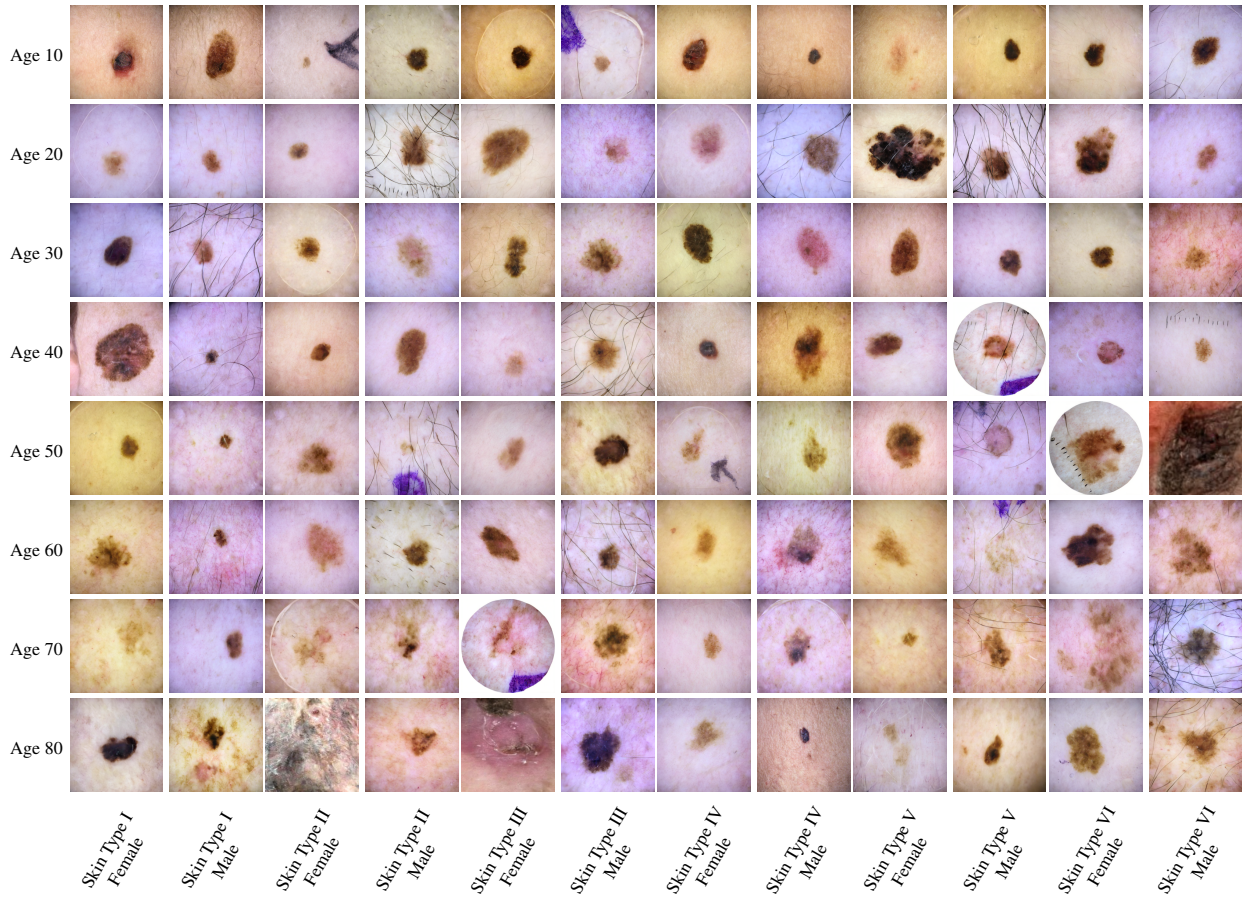


Figure 1. Synthetic melanoma images generated by our model with default sampling parameters (cfg=8, steps=200). Rows represent Fitzpatrick skin types (I–VI) combined with sex, and columns represent age groups (10–80). The grid demonstrates coverage of diverse demographic groups for fairness assessment.

semantics more reliably than general purpose CLIP models and is thus a good choice for general assessment. The zero-shot predictions were obtained using three prompt families corresponding to age, sex, and Fitzpatrick skin type. Age prompts covered the eight age groups and followed the template "A clinical close-up photograph of the {body_site} skin of a {age}-year-old melanoma patient." Sex prompts described the patient's sex using the format "A clinical close-up photograph of the {body_site} skin of a {sex} melanoma patient." Fitzpatrick skin type prompts covered types I through VI using the template "A clinical close-up of the skin of a melanoma patient with Fitzpatrick type {type} ({description of the skin type})." Overall, our synthetic images follow similar performance pattern across the age-sex-skin attributes, but our synthetic images are more recognizable as indicated by higher accuracies for all attribute groups. Interestingly, the performance across age groups is very imbalanced and varies a lot between real and synthetic showing that recognizing age from skin lesion images alone is difficult. Furthermore, both real and synthetic images are highly imbalances w.r.t to sex classification. Male accuracy is significantly lower than female accuracy, but our synthetic images are much better than real ones while reaching comparable performance for female. In terms of skin type classification, our synthetic images reach higher performance on all but skin type I, which is the best in real data. Finally, we visualize the CLIP score distributions by age, sex and skin type. Our synthetic images consistently lead to smoother distributions with higher mean values indicating better recognizability.

References

[1] Thorsten Kalb, Kaisar Kushibar, Celia Cintas, Karim Lekadir, Oliver Diaz, and Richard Osuala. Revisiting skin tone fairness in dermatological lesion classification. In *Workshop on Clinical Image-Based Procedures*, pages 246–255. Springer, 2023. 1

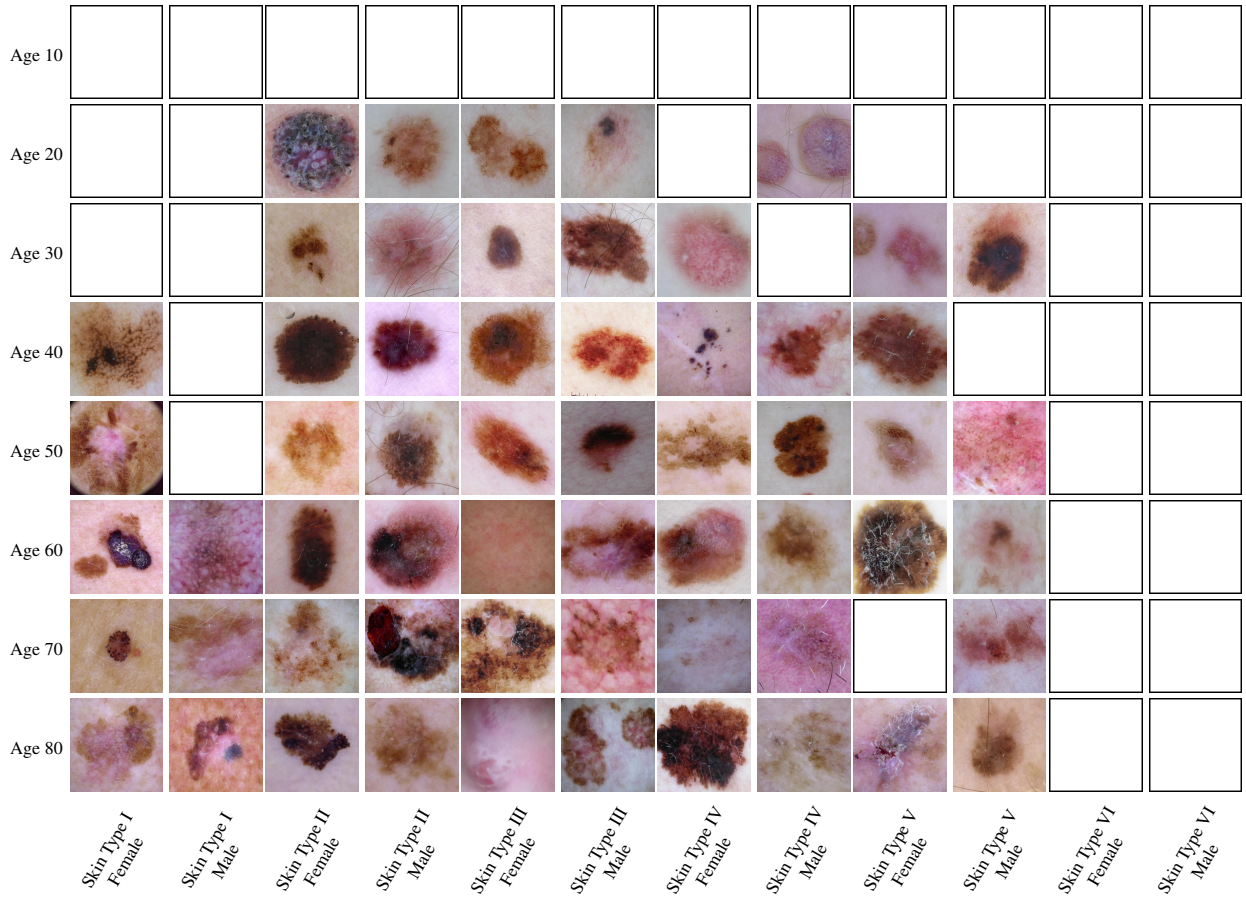


Figure 2. Real melanoma images from the MILK10k dataset [2]. Rows represent Fitzpatrick skin types (I–VI) combined with sex, and columns represent age groups (10–80). The grid demonstrates coverage of diverse demographic groups for fairness assessment. Many age–skin–sex combinations are missing from the real dataset, making fairness assessment difficult.

cfg \ steps	100	150	200	250	300	400
4.0	29.69	28.64	28.23	32.55	27.93	**27.77**
6.7	30.66	30.00	29.75	32.37	29.61	29.57
8.0	31.59	30.92	<u>30.75</u>	32.57	30.71	30.73
10.0	32.53	32.34	32.38	34.15	32.47	32.52
12.0	34.38	34.42	34.50	34.14	34.79	34.90

Table 1. FID averaged over age groups for each combination of classifier-free guidance (cfg) and diffusion steps showing that default setting (underlined) are not optimal settings (bold) when trained on skin lesions.

Age	Mean FID	Data ratio
10	40.56	0.35%
20	36.88	0.66%
30	36.05	3.13%
40	30.84	8.06%
50	26.08	11.57%
60	**25.74**	12.50%
70	27.60	9.37%
80	28.94	5.00%

Table 2. Data ratio and mean FID over all cfg and steps for each age group. Generation quality generally correlates with the available training data.

[2] MILK Study Team. Milk10k. ISIC Archive, 2025. 1, 3

[3] Schrasing Tong and Lalana Kagal. Investigating bias in image classification using model explanations. *arXiv preprint arXiv:2012.05463*, 2020. 1

[4] Yue Yang, Mona Gandhi, Yufei Wang, Yifan Wu, Michael S. Yao, Chris Callison-Burch, James C. Gee, and Mark Yatskar. A textbook remedy for domain shifts: Knowledge priors for medical image analysis. *arXiv preprint arXiv:2405.14839*, 2024. 1

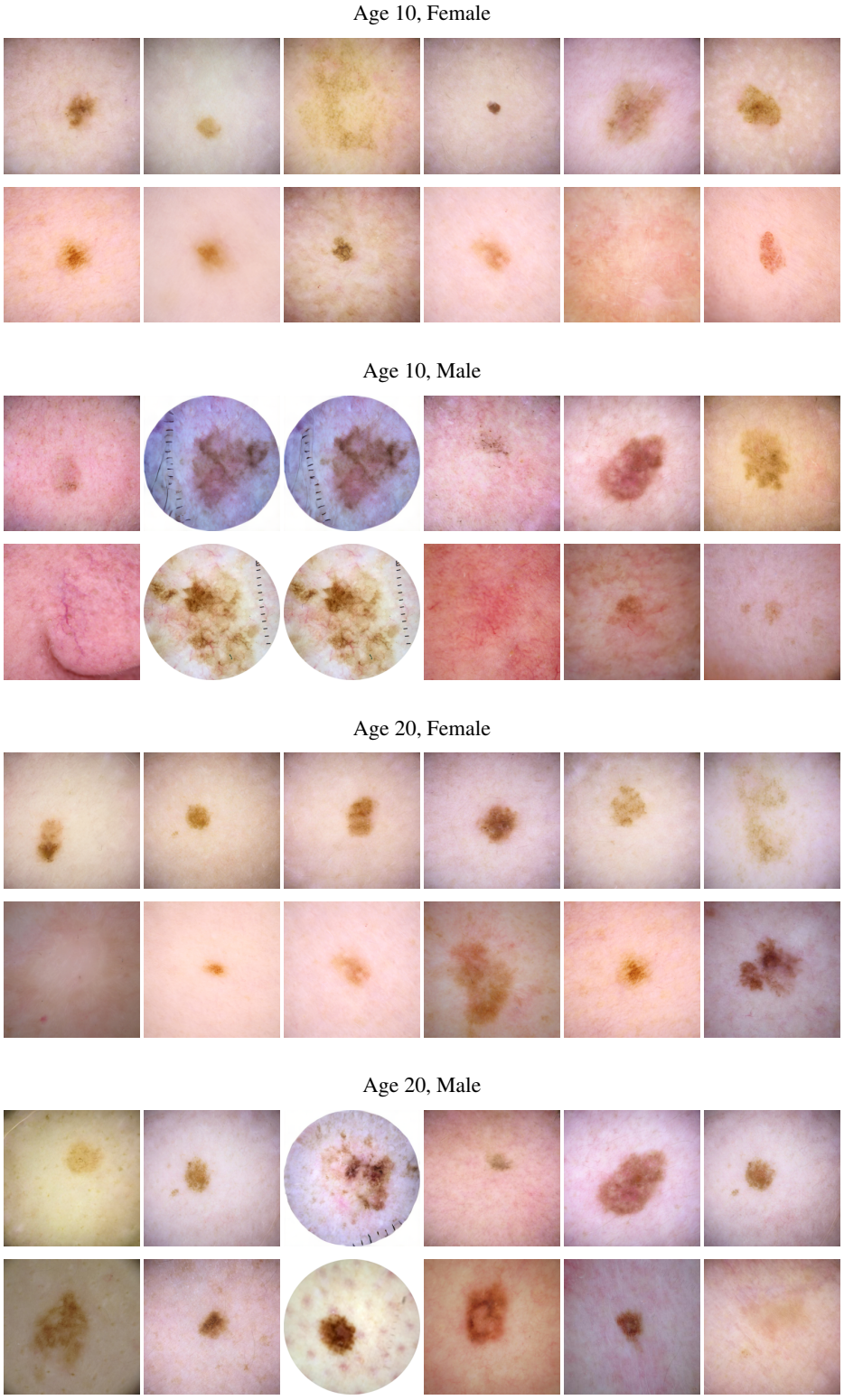


Figure 3. Synthetic samples (top row in each block) and nearest real neighbours (bottom row) for ages 10 and 20. Columns represent skin type I to VI.

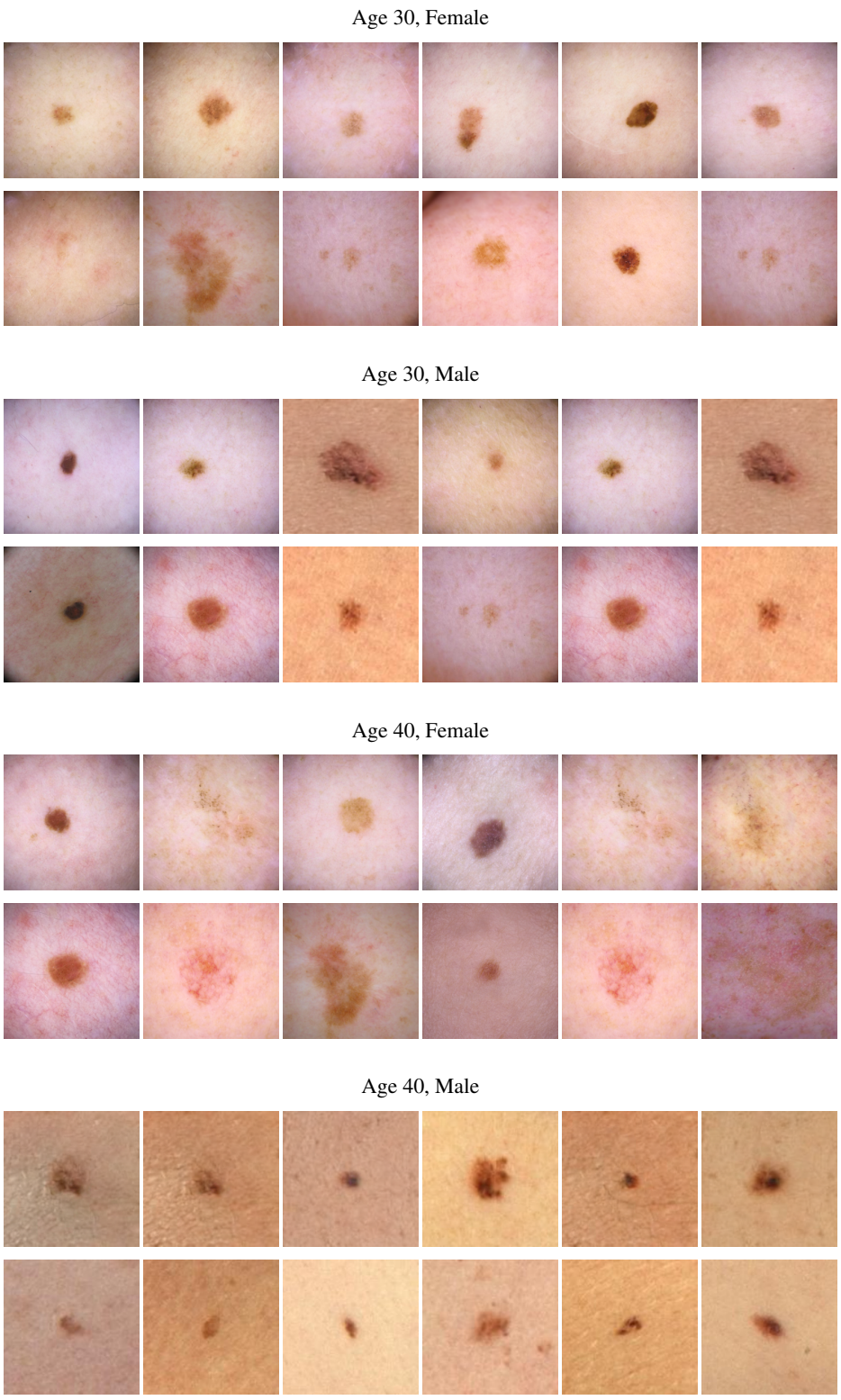


Figure 4. Synthetic samples (top row in each block) and nearest real neighbours (bottom row) for ages 30 and 40. Columns represent skin type I to VI.

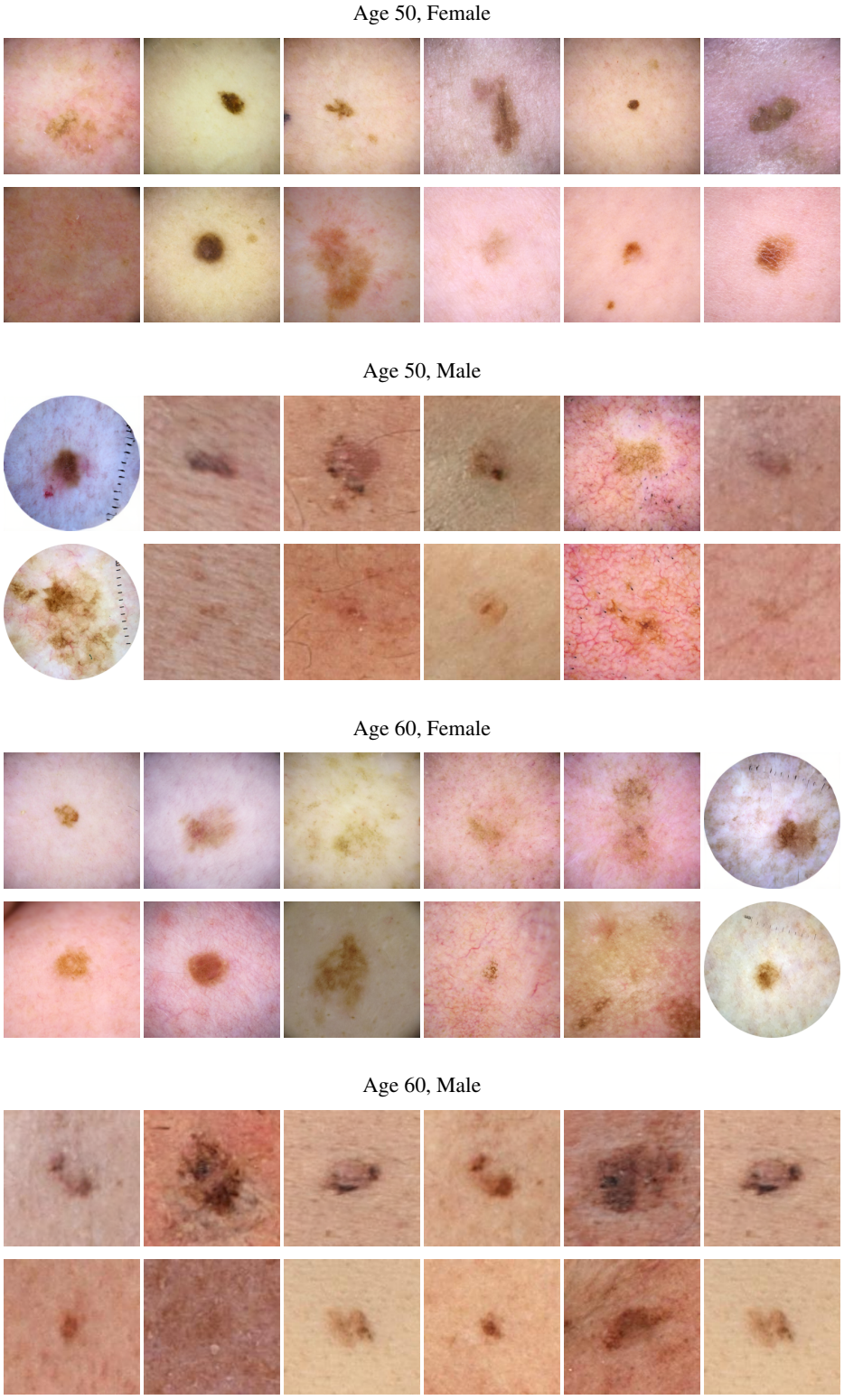


Figure 5. Synthetic samples (top row in each block) and nearest real neighbours (bottom row) for ages 50 and 60. Columns represent skin type I to VI.

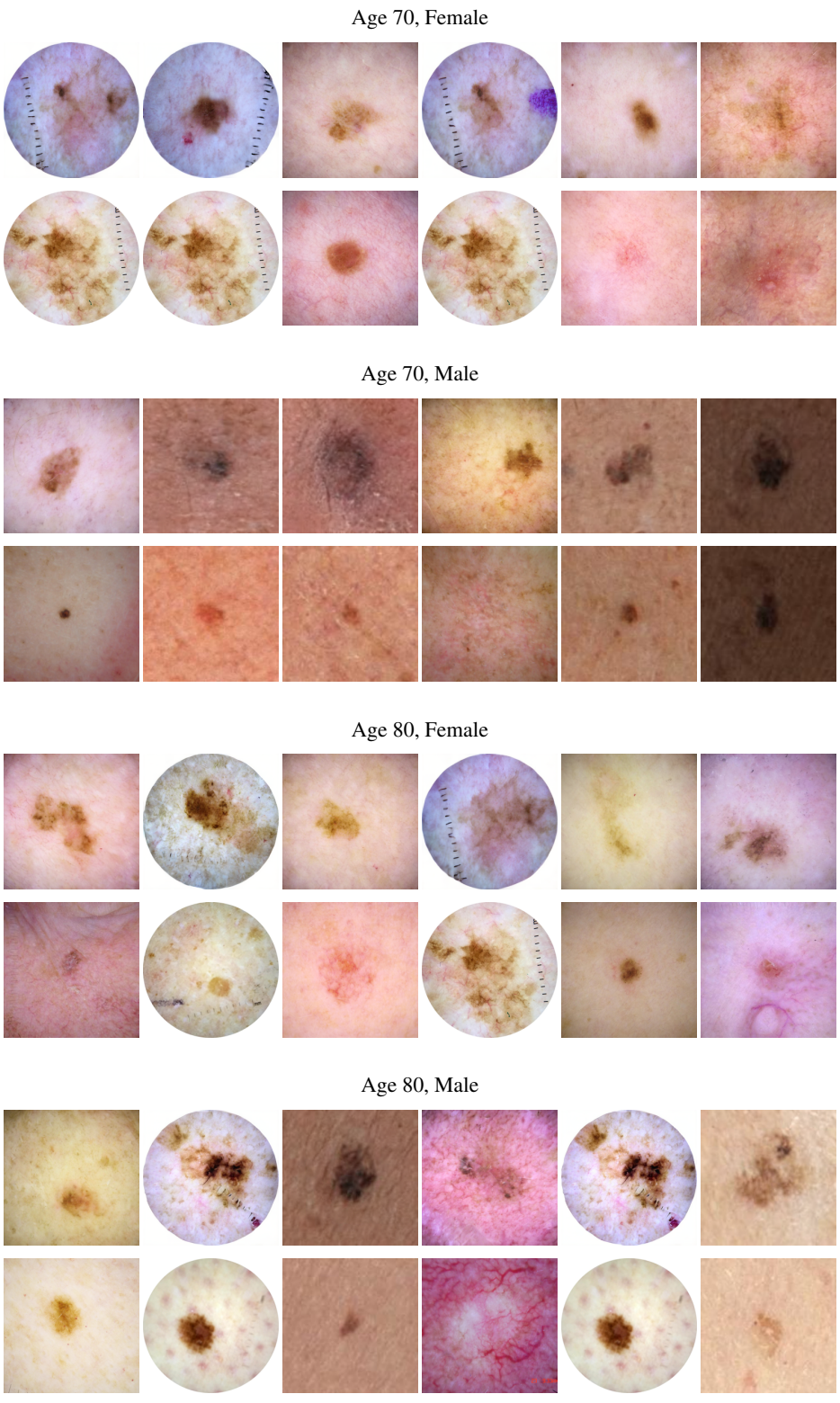
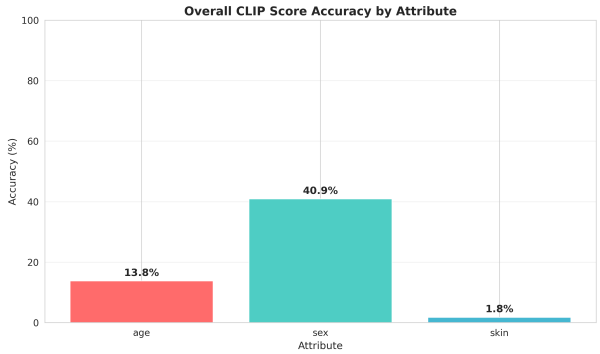
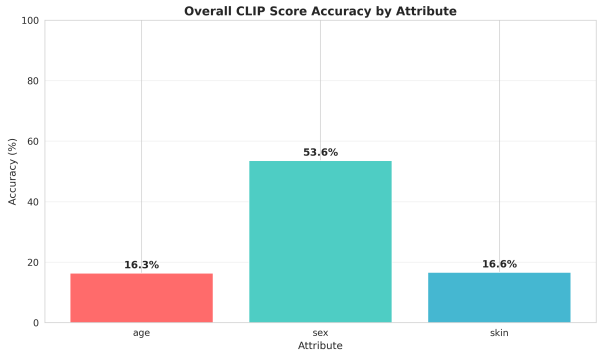


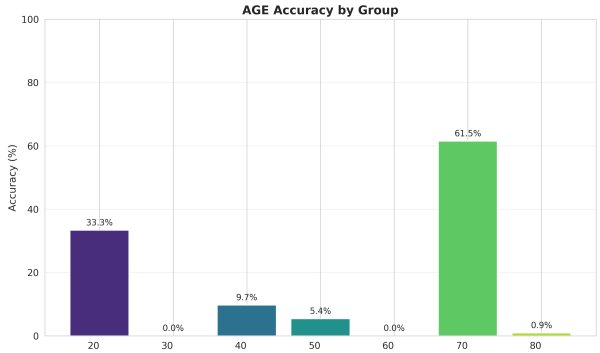
Figure 6. Synthetic samples (top row in each block) and nearest real neighbours (bottom row) for ages 70 and 80. Columns represent skin type I to VI.



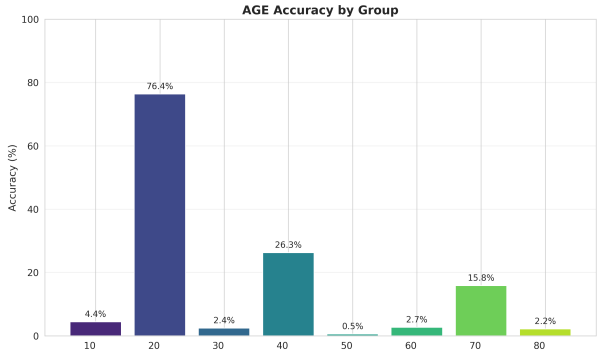
(a) Overall CLIP Score Accuracy (real)



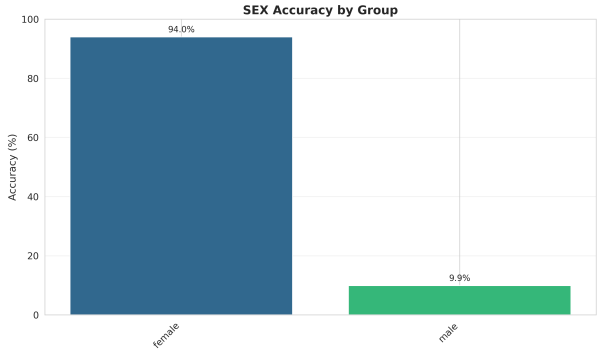
(b) Overall CLIP Score Accuracy (synthetic)



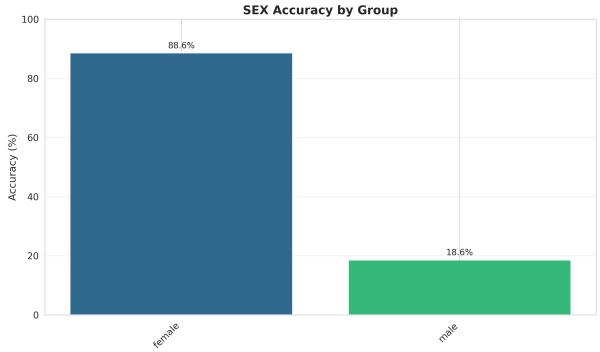
(c) AGE Accuracy (real)



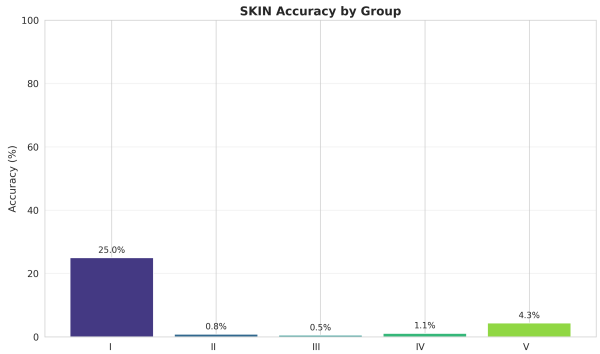
(d) AGE Accuracy (synthetic)



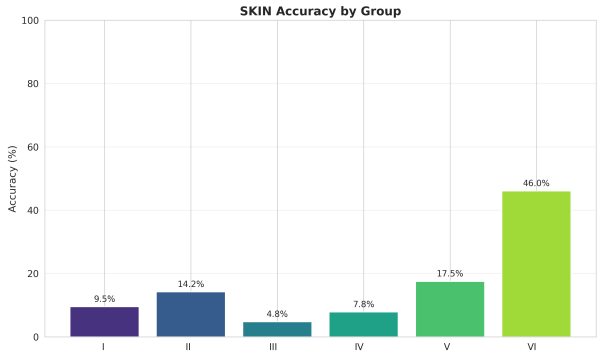
(e) SEX Accuracy (real)



(f) SEX Accuracy (synthetic)



(g) SKIN Accuracy (real)



(h) SKIN Accuracy (synthetic)

Figure 7. Comparison of CLIP-score classification accuracies across attributes.

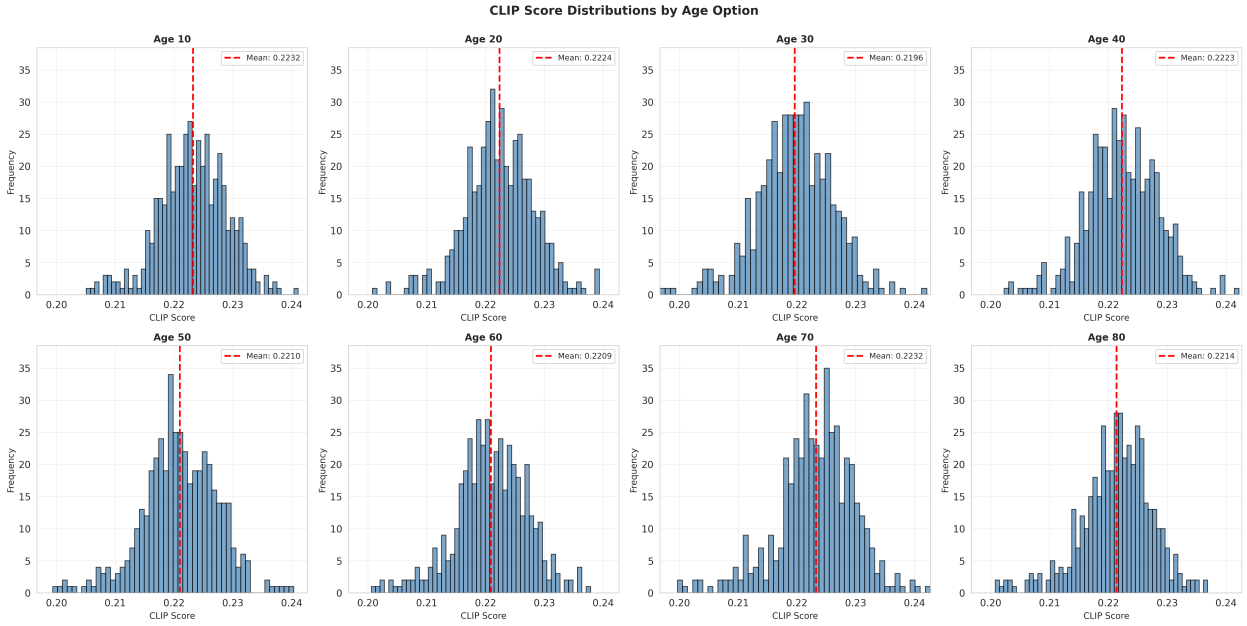


Figure 8. CLIP Score Distributions by Age (real)

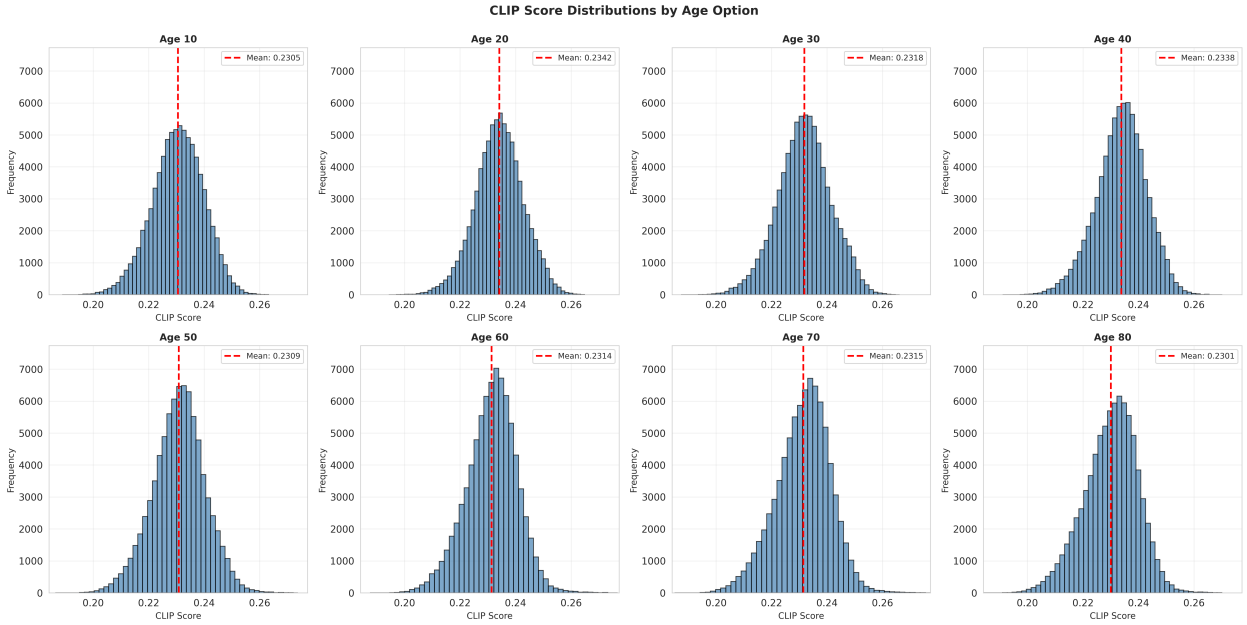


Figure 9. CLIP Score Distributions by Age (synthetic)

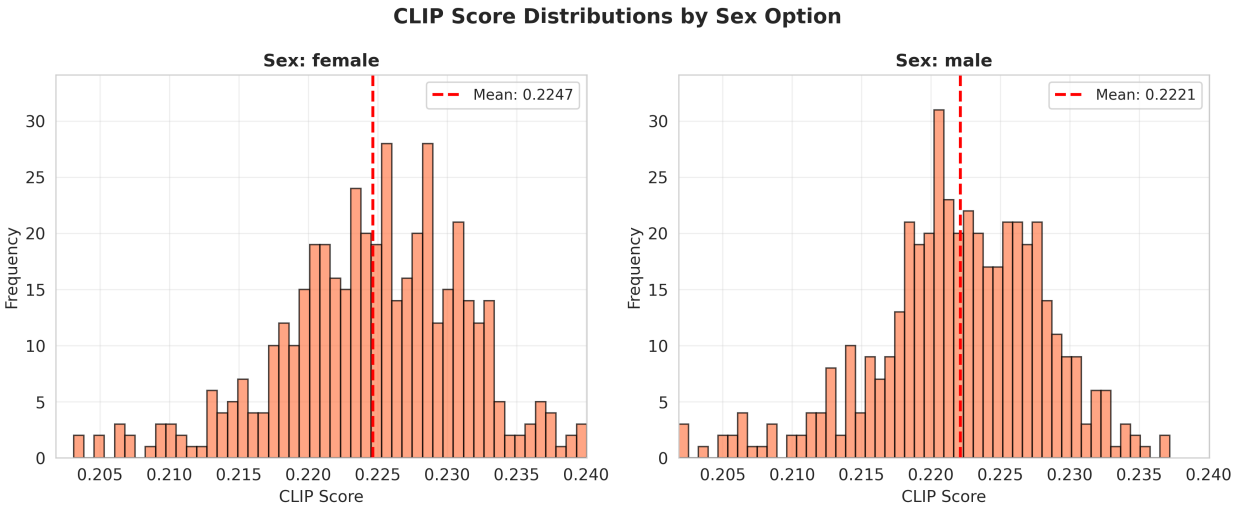


Figure 10. CLIP Score Distributions by Sex (real)

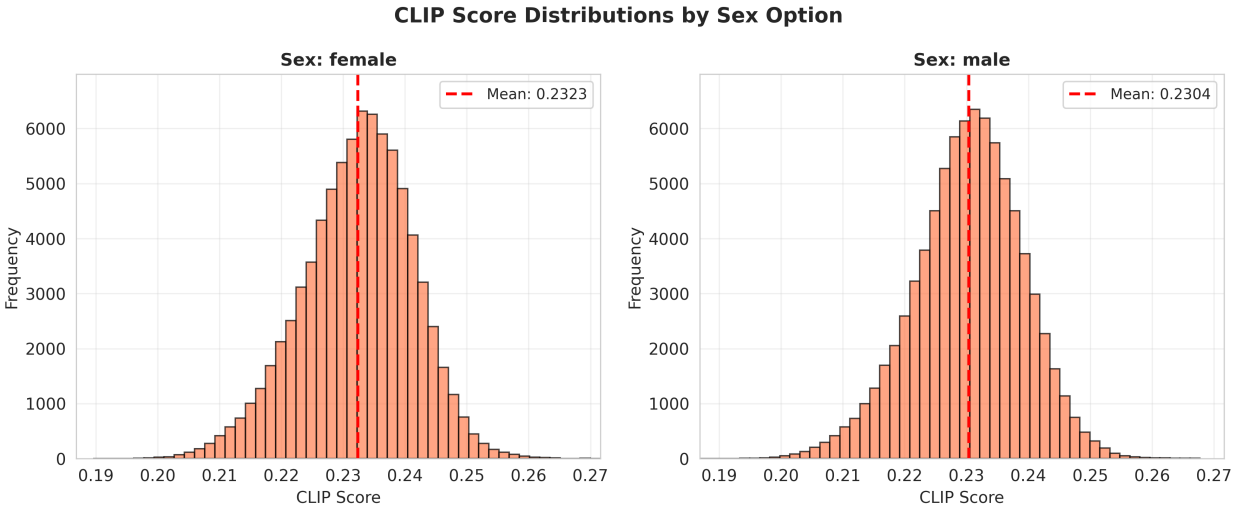


Figure 11. CLIP Score Distributions by Sex (synthetic)

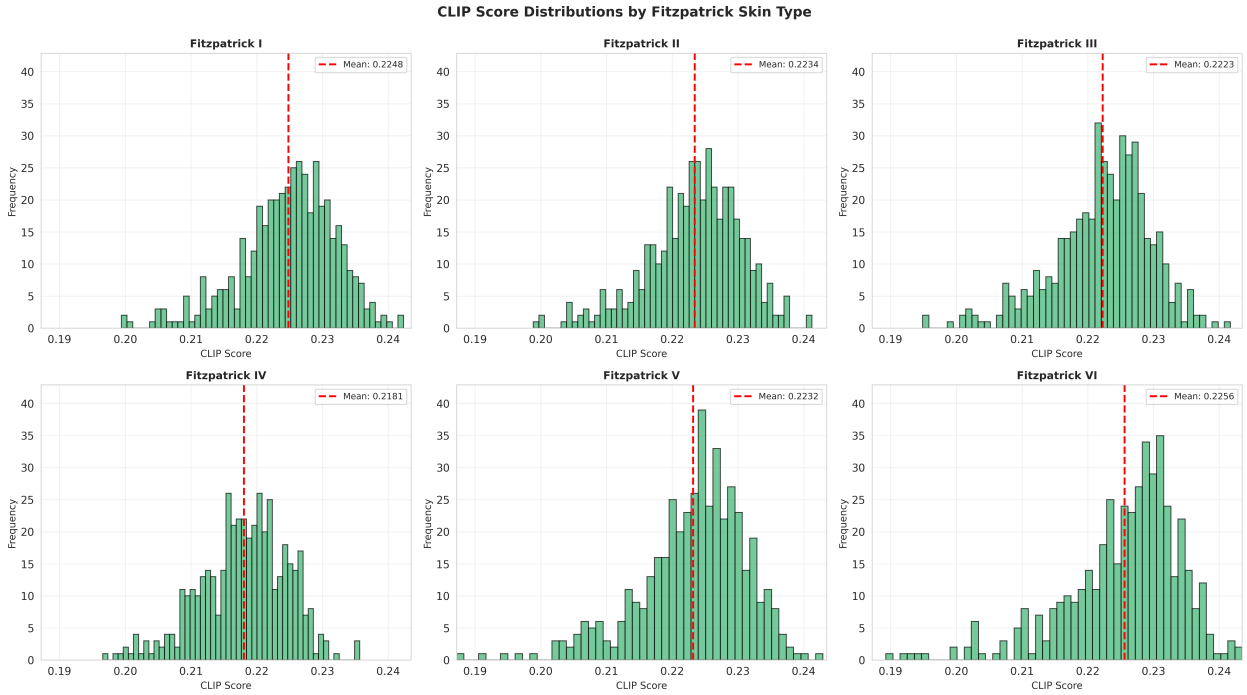


Figure 12. CLIP Score Distributions by Skin Type (real)

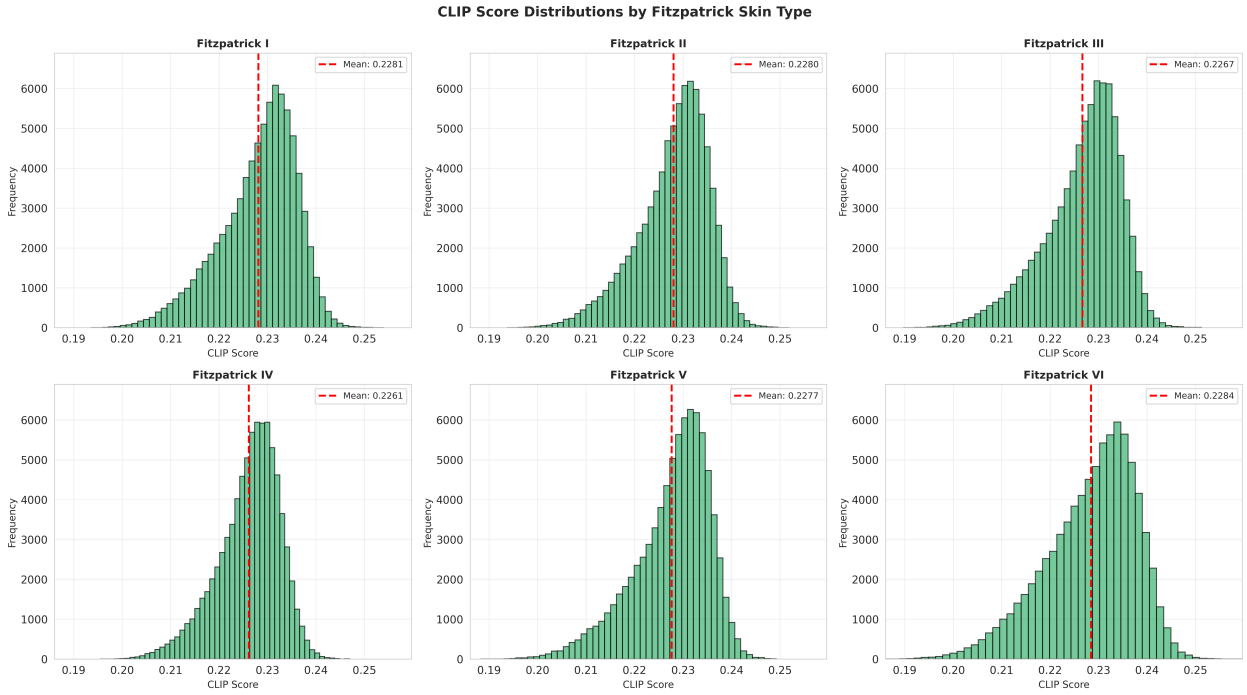


Figure 13. CLIP Score Distributions by Skin Type (synthetic)