

---

# Biased Attractiveness Bench (BAB): Image Reward Models Confuse Attractiveness with Realism

---

Harvey Mannering

Yaseen Mohammad  
Osman

Yeda Wang

Nehal Yasin

Annika Catulli

With  
Apart Research

## Abstract

Image reward models are used to finetune and improve diffusion models. Therefore any biases contained within these models can be exploited, leaving the generative model vulnerable to reward hacking. We explore how the attractiveness of faces impact the rewards assigned to an image, designing a new benchmark to measure the effect. In particular, we find that human preference models confuse attractive AI-generated faces with high quality, realistic images, assigning higher rewards for more attractive people. To our knowledge, we are the first to recognize this tendency in image reward models.

Code can be found at:

[https://github.com/harveymannering/AI\\_Manipulation\\_Hackathon](https://github.com/harveymannering/AI_Manipulation_Hackathon).

*Keywords: Image Rewards, Reward Hacking, Diffusion Models, Generative Models*

An exciting recent development in diffusion models is to use human feedback to improve generations. Following on from the success in language models, reinforcement learning with human feedback (RLHF) or supervised fine-tuning (SFT) can be used to refine an image generator to produce better quality images (Fan et al., 2023), (Clark et al., 2023). However, these techniques require reward models trained on human preference data. In language models the use of human

preference data has resulted in sycophancy bias, where LLMs will produce responses that correspond to a user's prior belief over what is true (Sharma et al., 2023). We look to explore where human preference models in images could lead to a similar effect.

Recent research shows that many image generation models exhibit different kinds of biases. Models tend to generate images that match mainstream beauty standards, often overriding prompts (Dinkar et al., 2025). Furthermore, many models are more likely to favour attractive faces when making unrelated decisions, for example in hiring, judicial sentencing and credit granting (Gulati et al., 2025). Some research also highlights that models that predict facial attractiveness contain ethnic biases (Hamel et al., 2025) which also extends to text generation in LLMs (Kamruzzaman., 2024).

Recent work in *human psychology* has found that attractive faces are less likely to be judged as artificially generated (Makowski et al., 2024). Inspired from this research we question whether this tendency replicates and causes issues for human preference models, where users label more attractive face images as looking “more realistic”. We therefore test if human preference models give higher rewards for more attractive faces, and find that, for some models, there are strong to medium correlations between the attractiveness of the faces and the rewards given. To our knowledge, we are the first to document that this bias exists in these models. Since human preference models are used to finetune diffusion models, either with Reinforcement Learning with Human Feedback (RLHF) or Supervised fine-tuning (SFT), this bias leaves diffusion models vulnerable to reward hacking, where learning to maximize human preference results in identifying attractive faces as real, and non-attractive as AI-generated.

Our contributions are:

- We develop a synthetic facial dataset annotated for attractiveness, gender, and age.
- We show that diffusion models tend to associate attractiveness as being female.
- We provide empirical evidence that reward models classify attractive faces as more realistic and less attractive faces as more likely AI-generated.

# 1. Methods

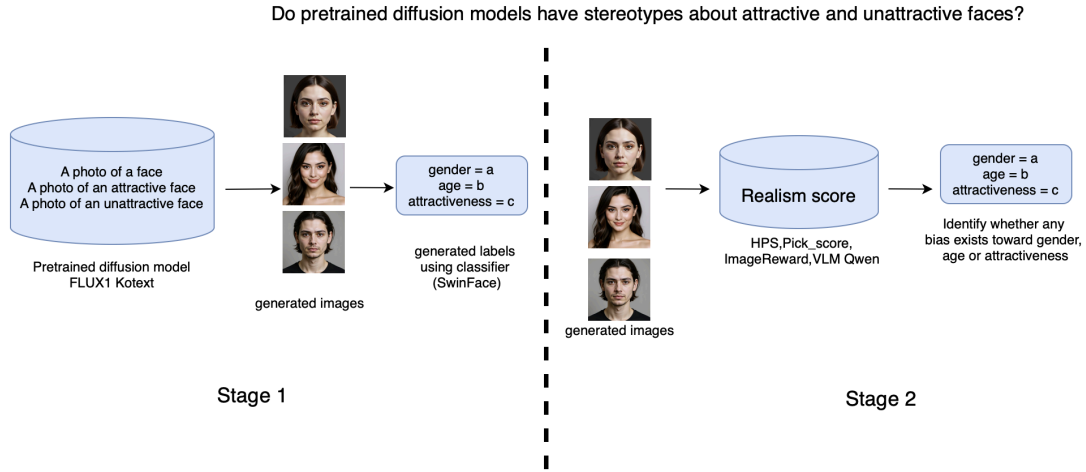


Figure 1: Overview of Pipeline: Stage1: We generate images from diffusion model and use a classifier to annotate the images. Stage2: The generated images are evaluated by a reward model that assigns each image a realism score. In the final step, we assess whether the reward models exhibit bias with respect to attributes such as gender, age, or attractiveness when predicting image realism.

To measure facial beauty bias in human preference models we create a benchmark of 300 synthetically generated images. We use FLUX.1 Kontext to generate images using the prompts:

- "a portrait photo of an attractive face"
- "a portrait photo of a face"
- "a portrait photo of an unattractive face"

In our prompts we specify “attractiveness” because “beauty” is culturally associated with women. For each prompt we generated 100 images using 40 inference steps. We hypothesize that various reward/human preference models (used in RLHF for diffusion models) will have picked up the bias from humans that more attractive faces look more realistic.

Our benchmark also includes labels for the estimated age and gender of each image obtained using the SwinFace model (Qin et al., 2023). Interestingly, we find that FLUX.1 Kontext is significantly more likely to generate a female face if the word attractive is included in the prompt.

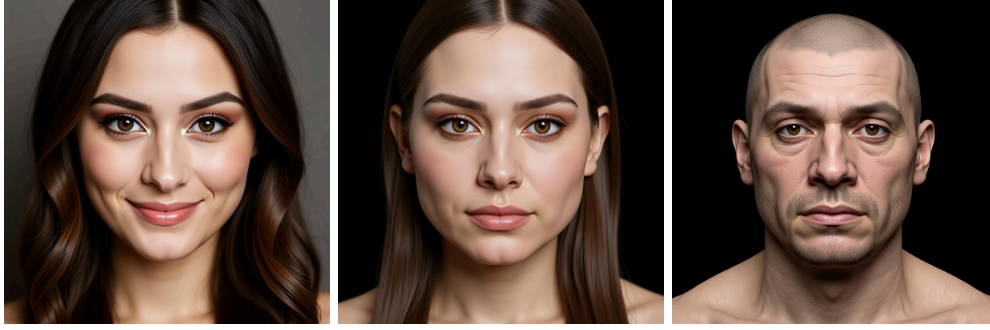


Figure 2 - Example images from our dataset generated using FLUX.1 Kontext. These example images used the same starting seed but different prompts.

## 2. Results

We calculate the score for all 300 images in our dataset for four image reward models (including three human preference models): ImageReward (Xu et al., 2023), PickScore (Kirstain et al., 2023), HPSv2 (Wu et al., 2023), and Qwen3-VL-8B-Instruct (Bai et al., 2025). These models are trained to predict which AI generated images humans prefer and are often used as a proxy for image quality.

For Qwen3-VL-8B-Instruct, we followed a generative procedure to evaluate the model. Specifically, we prompted the model to decide how likely the image is AI-generated on a scale from 1 to 10, with 10 being highly likely and 1 being unlikely. The prompt template we have used can be found on the Appendix. Finally, we normalised the output following a similar normalisation procedure used by Rajpurkar et al. (Rajpurkar et al., 2018). Several in-depth figures can be found on the Appendix. Evaluation setup for ImageReward, HPSv2, and PickScore can be found in the Appendix.

In table 1 we calculate the correlation coefficient between the between human preference scores and the estimated facial attractiveness of each (predicted using SwinFace).

| Model                | Correlation Coefficient |
|----------------------|-------------------------|
| ImageReward          | -0.0701                 |
| PickScore            | 0.6141                  |
| HPSv2                | 0.5961                  |
| Qwen3-VL-8B-Instruct | 0.2671                  |

Table 1 - Correlations between perceived facial attractiveness in our generated images and the reward assigned to each image.

We find that for PickScore and HPSv2, more attractive faces have a strong to medium correlation with the human preference score, suggesting that the psychological tendency to see more attractive faces as more realistic may have been distilled into current human preference models for vision.

### 3. Discussion and Conclusion

In this work, we create a benchmark of 300 AI-generated face images with attractiveness labels. These can then be used to probe reward models to detect if there is a bias towards attractive faces. We find that image reward models, and in particular human preference models, which are often used as a proxy for image quality, label more attractive facial images as better quality. To the best of our knowledge we are the first to notice this tendency in human preference models.

Since these models are used to finetune diffusion models, via either RLHF or SFT, this has concerning implications for reward hacking. PickScore and HPSv2 are intended to improve image quality and text-image alignment, however, our results suggest that using these models as rewards during fine-tuning risks improving facial attractiveness at the expense of true alignment.

Studies suggest that the dominance of attractive faces on social media platforms and in the media has led to an increase in occurrences of body dysmorphic disorder (Laughter et al., 2023). This phenomena would only increase with the majority of artificially generated faces being attractive.

Additionally, generative models have been shown to encode westernised beauty conventions (Stacy, 2025). This adversely affects young people of different ethnicities who may grow up believing that whiteness is the standard for beauty.

### 4. Future Work

Our study is based completely on AI-generated images as it requires special ethical approvals to use human faces data. There are currently real images datasets

available (e.g. CelebA, BFFHQ etc) that can be further used to get more evidence for this study. We strongly believe that there is a bias that current models find attractive faces to be more realistic. This study can be continued to perform further experiments on real-world datasets. Lastly, we plan to expand our investigation of VLMs to other model architectures and sizes.

## 5. References

- Bai, Shuai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibong Song, Kai Dang et al. (2025). Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.
- Clark, Kevin, Paul Vicol, Kevin Swersky, and David J. Fleet. (2023). Directly fine-tuning diffusion models on differentiable rewards. *arXiv preprint arXiv:2309.17400*
- Dinkar, Tanvi, Jiang Aiqi, Gavin Abercrombie, Ioannis Konstas. (2025). Erasing 'Ugly' from the Internet: Propagation of the Beauty Myth in Text-Image Models. 10.48550/arXiv.2511.00749.
- Fan, Ying, Olivia Watkins, Yuqing Du, Hao Liu, Moonkyung Ryu, Craig Boutilier, Pieter Abbeel, Mohammad Ghavamzadeh, Kangwook Lee, and Kimin Lee. (2023). Dpok: Reinforcement learning for fine-tuning text-to-image diffusion models. *Advances in Neural Information Processing Systems* 36 : 79858-79885.
- Gulati, Aditya, Moreno D'Inca, Nicu Sebe, Bruno Lepri, Nuria Oliver. (2025). Beauty and the Bias: Exploring the Impact of Attractiveness on Multimodal Large Language Models. 10.48550/arXiv.2504.16104.
- Hamel, Chandon, Mike Busch. (2025). Analysis of Bias in Deep Learning Facial Beauty Regressors. 10.48550/arXiv.2509.24138.
- Kamruzzaman, Mahammed, Md. Minul Shovon, Gene Kim. (2024). Investigating Subtler Biases in LLMs: Ageism, Beauty, Institutional, and Nationality Bias in Generative Models. 10.48550/arXiv.2309.08902.
- Kirstain, Yuval, Adam Polyak, Uriel Singer, Shahbuland Matiana, Joe Penna, and Omer Levy. (2023). Pick-a-pic: An open dataset of user preferences for text-to-image generation. *Advances in neural information processing systems* 36: 36652-36663.
- Laughter, Melissa & Anderson, Jaclyn & B C Maymone, Mayra & Kroumpouzou, George. (2023). Psychology of Aesthetics: Beauty, Social Media, and Body Dysmorphic Disorder. *Clinics in Dermatology*. 41. 10.1016/j.clindermatol.2023.03.002.
- Makowski, Dominique, Te An Shu, Ana Neves, Stephanie Kirk, Ngoi Zi Liang, Panos Mavros, Chen SH. (2024). Too beautiful to be fake: Attractive faces are less

likely to be judged as artificially generated. *Acta psychologica*. 252. 104670. 10.1016/j.actpsy.2024.104670.

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know What You Don't Know: Unanswerable Questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics

Qin, Lixiong, Mei Wang, Chao Deng, Ke Wang, Xi Chen, Jiani Hu, and Weihong Deng. (2023). Swinface: a multi-task transformer for face recognition, expression recognition, age estimation and attribute estimation. *IEEE Transactions on Circuits and Systems for Video Technology* 34, no. 4: 2223-2234.

Sharma, Mrinank, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Newton Cheng et al. (2023) Towards understanding sycophancy in language models. *arXiv preprint arXiv:2310.13548*.

Stacy, H. R. (2025). The Representation of Feminine Beauty in Generative Artificial Intelligence Models.

Wu, Xiaoshi, Yiming Hao, Keqiang Sun, Yixiong Chen, Feng Zhu, Rui Zhao, and Hongsheng Li. (2023). Human preference score v2: A solid benchmark for evaluating human preferences of text-to-image synthesis. *arXiv preprint arXiv:2306.09341*

Xu, Jiazheng, Xiao Liu, Yuchen Wu, Yuxuan Tong, Qinkai Li, Ming Ding, Jie Tang, and Yuxiao Dong. (2023). Imagereward: Learning and evaluating human preferences for text-to-image generation. *Advances in Neural Information Processing Systems* 36: 15903-15935.

Yu, Cai, Shan Jia, Jiahe Tian, Jiao Dai, Jizhong Han, Siwei Lyu. 2025. Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV) Workshops,, pp. 1553-1563

## 6. Appendix

### 6.1. Security Considerations

Below we have listed the **security considerations** for this project:

- Limitations - While we have found several strong correlations suggesting attractiveness bias, correlation does not mean causation. Further analysis is needed to determine if attractive faces have a causal link to the higher rewards.
- Dual-use risks - We do not believe our method can be used to train better manipulators.
- Responsible disclosure recommendations - The project has avoided the use of any personal or sensitive data.

- Ethical considerations in your approach - We intentionally used AI-generated images to avoid ethical concerns around real face image datasets.
- Suggestions for future improvements - Examine reward hacking on diffusion models using these reward model. Do these biases also exist when these reward models are run on real image datasets?

## 6.2. Human Preference Model Evaluation Setup

Prompts used for ImageReward, HPSv2, and PickScore were kept constant as "a portrait photo of a face". In the main text attractiveness labels were obtained via the SwinFace classifier, however we have also performed analysis with the using labels obtained from the generation prompt (i.e. images generated with the prompt "a photo of an attractive face" are labelled as attractive, images generated with the prompt "a photo of an unattractive face" were labelled as unattractive). We found similar correlations using these labels as with the classifier derived labels.

For the ImageReward, HPSv2, and PickScore results in the main body, we include the following plots from which the correlation coefficients were derived. The y-axis shows the attractiveness labels for the images (higher for more attractive) whereas the x-axis shows the reward assigned to the image.

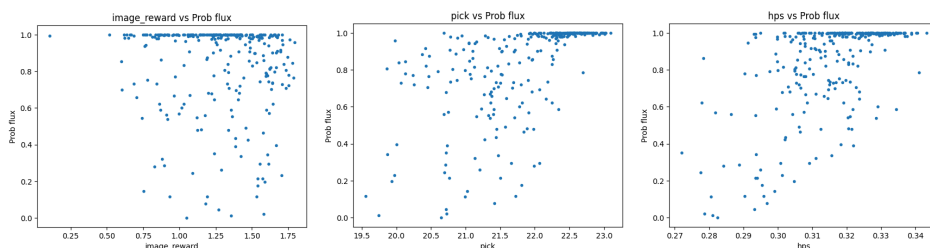


Figure 3 - Correlation plots from human preference models ImageReward, PickScore and HPSv2.

## 6.3. Qwen3-VL-8B Prompt

Below we share the prompt used during Qwen3-VL-8B inference. This was inspired by commonly used prompting styles for this model as well as similar tasks (Bai et al., 2025; Cai et al., 2025). The prompt is added following the input image to evaluate.

This is an image of a human. Measure how likely it is AI-generated on a scale between 1 to 10, with 1 being highly unlikely and 10 being highly likely.

## 6.3. Additional Qwen3-VL-8B Results

We have observed that Qwen3-VL-8B, when used to measure how likely an image is AI-generated, has a weak correlation with SwinFace attractiveness classifications. Moreover, when using the labels in our benchmark that were used to generate the data, a much higher correlation is seen, suggesting a closer connection between Qwen3-VL-8B and FLUX.1 Kontext. Additionally, most of Qwen3-VL-8B predictions ranged between 7-9 suggesting the need for a thorough

investigation including prompts-refinement. Lastly, we plan to conduct a more rigorous and comprehensive VLM investigation in the future.

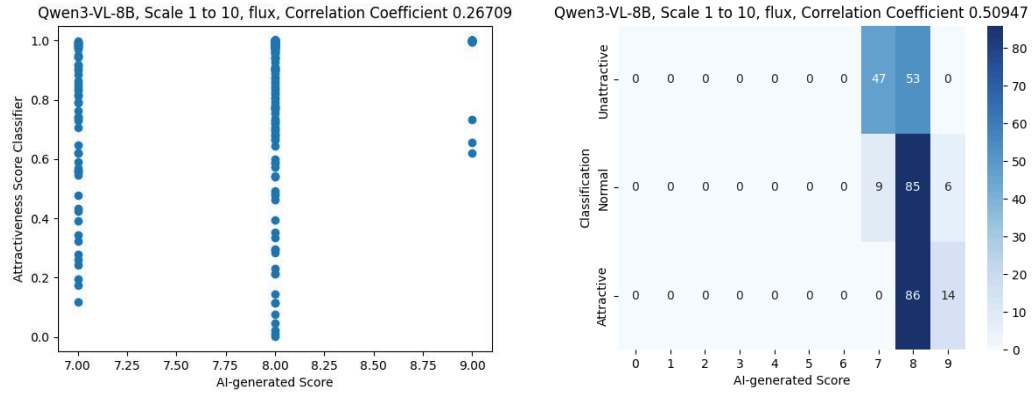


Figure 4 - Shows Qwen3-VL-8B results when evaluated on our benchmark. (Left) Shows a weak correlation appearing between Qwen3-VL-8B classification of AI-generated images and SwinFace classification of attractiveness. (Right) A much higher correlation is seen when using our benchmark labels.