BIAS-101: Identifying and Mitigating Bias in UCF-101 Action Dataset

Adam Bawatneh
University of Central Florida
Adam.Bawatneh@ucf.edu

Colin Houde
University of Central Florida
colin.houde@ucf.edu

Scott Spicer
University of Central Florida
scott.spicer@ucf.edu

Zengyan Wang
University of Central Florida
zengyan.wang@ucf.edu

Mubarak Shah*
University of Central Florida
shah@crcv.ucf.edu

Abstract

When training a deep neural network to perform any task, robustness and generalization are essential. In order to train a successful and robust classifier, the model must be trained on all possible use cases to generalize and perform well regardless of what it is being asked to do. As we shift towards larger, more robust models, it is important to train the model on all possible use cases and scenarios that can be found within the real world. This becomes increasingly difficult to do when using datasets that contain bias. Bias within datasets ultimately hinder the model to generalize well and lowers the ceiling of maximum possible robustness. When utilizing datasets that have bias within them, the models also inherit that bias and relay it forward creating unfair and inaccurate predictions. Bias mitigation techniques can be deployed but are often expensive and do not cure the core problem but apply more of an on-the-top fix. If we train on a dataset that contains minimal bias, the model will then in turn contain minimal bias itself. In this study, we take a closer look at one specific action recognition video dataset, UCF-101, to discover and address the bias within, in order to mitigate and rebalance the dataset as a whole. In doing so, we debias the dataset and create a more generalized action recognition video dataset for future models to utilize and, in turn, become more robust themselves.

1 Introduction

The UCF-101 [16] dataset is a widely recognized benchmark in the field of computer vision. It was one of the first datasets to specialize in action recognition in videos and is still used to examine model performance within the subfield. UCF-101 is known for its 101 action categories, covering a broad spectrum of human activities including but not limited to sports and typical daily tasks. Spanning over 13k videos and 27 hours of video data, the dataset is unique as it offers a 'in the wild' approach that makes it a solid pillar for benchmarks and testing. Although the dataset contains ample data across multiple domains, there is still an inherent bias within. In order to maintain an honest and fair benchmark, we must discover and address the bias so we can rebalance and mitigate it. There is no beneficial reason for a dataset to contain any form of bias. Bias within datasets creates a plethora of problems. Not only does it create inaccurate and unreliable artificial intelligent systems, but it can also cause harm to people and businesses. When training a classifier or a model on a dataset that contains bias, the model may develop a prejudiced decision-making process which can lead to unfair and discriminatory outcomes. For instance, in computer vision applications like facial

*Faculty advisor. Contact: shah@crcv.ucf.edu

39th Conference on Neural Information Processing Systems (NeurIPS 2025).

recognition, a biased dataset that over-represents a particular ethnicity can result in higher error rates for underrepresented groups. This not only undermines the reliability of the system but can also perpetuate social inequities. Bias within datasets come in many forms. While it is exponentially difficult to address and mitigate all bias, in this work we aim to address and rebalance the Selection, or Representation Bias within UCF-101. Selection bias occurs when the dataset does not fully represent the diversity of the real-world population or scenarios the model is intended to handle. This happens when certain groups, conditions, or environments are over- or under-represented. For example, a facial recognition dataset composed mostly of light-skinned individuals may cause the model to perform poorly on darker-skinned faces, resulting in higher error rates for those groups. This creates a model that struggles to generalize new data, leading to a less robust and less trustworthy model. We approach this problem through a systematic hand-crafted approach. We first determine the demographic bias types to identify the niche biases within the dataset. We then generate labels and perform statistical analysis to show us where the bias is prevalent and the amount of new data we need in order to rebalance and debias the dataset. We finally repeat the process to make sure there is statistically no bias left within the demographic groups in the dataset. While it may be impossible to address and rebalance all forms of bias within a dataset, using this process we work to vastly mitigate it and create a far more fair, well rounded action recognition video dataset.

Figure 1: The 101 action categories found in the UCF-101 dataset. This image illustrates the wide diversity of actions represented.

2 Related Work

There has been a large focus on ensuring that models that are created are unbiased. This has led to the realization that because models learn from datasets, they will likely output solutions that reflect the bias in those datasets. It has been shown that the less biased dataset the less biased the model will be [13]. Further research into datasets has shown that they commonly contain biases both known and unknown [17].

Different strategies have been employed to audit and detect these biases. The authors of this paper [21] use VLMs to caption images to detect known and unknown biases to then be further debiased. Another very intuitive and common technique is to employ clustering algorithms to cluster the data in groups to reveal bias [15]. The REVISE tool released in 2021 [18] aids in investigating visual biases across three different dimensions object, personal, geography based. One of the earliest ideas to mitigate bias came because of performance issues from datasets that were largely unbalanced. In this paper [2], the authors proposed a now well-known technique SMOTE. The idea is to oversample the underrepresented classes in the training data to improve the generalization of the model. The success of this technique led to other techniques enhancing it. An example of this is an approach that combines k-means clustering with SMOTE to overcome class imbalances [3]. Another paper [9] proposes GenSample, which employs a genetic algorithm to determine the optimal rate of oversampling. While these techniques have seen success, they suffer from needing to rely on the data you have. If the gap between the number of samples in the majority class and the minority class is too great the results will be limited.

Similar to oversampling, the idea of having more of the less represented data to solve the problem of biased datasets is solved with synthetic data. Synthetic data around a while. In fact, in this paper back in 2008 [5] the authors propose Adaptive synthetic sampling (ADASYN). It uses a weighted distribution for the minority of classes based on how difficult they are to learn. Then this weighted distribution is used to determine how much of each minority of class to generate. In a more recent paper, the authors use stable diffusion with cross attention to create sets of images with the differences isolated to the social attribute while keeping the rest of the image the same [6]. The authors of this paper [8] generated synthetic data using game engines to simulate scenarios and extracting data from the various scenarios, aiding in obtaining hard to get data. A common technique is to use generative adversarial networks (GANs). This paper [4] uses a conditional Wasserstein GAN that can effectively model tabular datasets with numerical and categorical variables and generate synthetic data for underrepresented classes. These approaches suffer from potentially noisy generation that are not accurate to real world data. This could hurt the performance of the model as it could associate the underrepresented classes with features hidden in artificially generated data, both human visible and not.

Some work has been done to train models specifically to dataset bias mitigation. In fact, the authors of this paper [10] created a dataset of $108,501$ images of various faces for the sole purpose of aiding in the training of models for bias detection, which in this case would be racial bias. The authors of this paper [20] use semi-supervision to improve the bias in pseudo-labels of the dataset. Another approach uses meta learning with an inner loop and an outer loop which plays an adversarial game against each other [1]. The goal is to find a split for the data so that the predictors learned on the training split cannot generalize to the testing split. The problem with these types of approaches is that you would need a separate dataset apart from the one you are trying to debias. In fact, you would need one for each type of bias you have so your debiasing model has experience with that type of bias. Also, how exactly a model can debias a dataset is less straightforward than previously mentioned methods.

3 Methodology

This section details our approach in three parts:

3.1 Bias Identification

We begin by identifying the primary demographic bias types present in the UCF-101 dataset. Rather than assuming predefined categories, we manually reviewed the dataset to observe recurring identity-related imbalances. To do this efficiently, we examined only the first frame of each sample, as samples

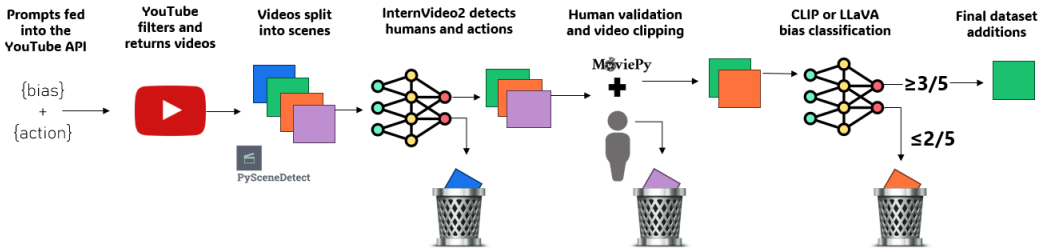

Figure 2: Our proposed workflow

within the same action category are often sequential clips from the same source video. This approach allowed us to reliably identify common bias types without redundant visual analysis. Through this process, we determined five major bias types: gender, race, age, hair color, and body type.

Each attribute was discretized into fixed categorical labels. Gender includes male and female. Race includes white, black, middle eastern, indian, asian, and hispanic. Age includes kid, young, and old. Hair color includes blonde, red, white, and black. Body type includes thin and wide.

To annotate each video with demographic labels, we performed zero-shot classification using CLIP [14] and LLaVA-OneVision [12]. Both models were applied to all five demographic bias types for every video in the dataset. For each video, five frames were randomly sampled using the strategy proposed in [7], with a fixed random seed to ensure reproducibility. The choice of five frames was guided by computational efficiency, as it balances speed and robustness while enabling majority voting. An odd number was selected specifically to minimize ties in label aggregation.

CLIP was used for attributes that lend themselves to retrieval-style classification via prompt matching. For instance, gender classification prompts took the form: "A photo of a male doing {activity}" versus "A photo of a female doing {activity}", where the activity is extracted from the filename. LLaVA-OneVision was used for open-ended attributes like age or body type, with prompts such as: "What is the most likely body type of the person doing {activity}? Choose from thin or wide."

We adopted both CLIP and LLaVA to assess the comparative effectiveness of retrieval-based versus generative vision-language models. While CLIP is typically more efficient, we evaluated whether generative outputs from LLaVA-OneVision offered improved accuracy on nuanced attributes.

Label decisions for each attribute were determined via majority voting across the five sampled frames. If no single category received a majority (e.g., a tie of two votes each for two classes), the label was marked as uncertain. This mechanism improved label reliability and reduced noise [11].

To further improve consistency, we implemented intra-video smoothing. Since UCF-101 videos are often split into short clips from a common source, we grouped clips from the same video and aggregated their predictions. The dominant class was assigned across the group. For example, if one source video produced labels: white, white, white, white, black, the final race label was white. This procedure reduced intra-source label variance and improved consistency across the dataset.

This structured annotation process—bias type identification, multi-model zero-shot classification, majority voting, and intra-video smoothing—enabled scalable and reliable demographic labeling for downstream analysis.

3.2 Statistical Analysis

To quantify the extent of bias in the UCF101 dataset, we employed both Chi-Square tests and a normalized Dominance Ratio to comprehensively assess demographic imbalance across different action categories. The Chi-Square test is a statistical method used to determine whether there is a significant difference between the observed and expected distributions of categorical data. It calculates the difference using the formula:

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

Where O represents the observed frequency of a particular demographic group in a given action category, and E is the expected frequency, assuming an unbiased distribution. A higher Chi-Square value indicates a greater deviation from the expected balance, suggesting the presence of bias in the dataset. While the Chi-Square test provides statistical validation of bias, it does not directly measure how much a particular class dominates over others. To address this, we introduce the Normalized Dominance Ratio, which quantifies the degree to which one demographic class is overrepresented compared to all others within an action category. The Dominance Ratio is calculated as:

$$DR = \frac{M}{(M+O)N}$$

Where M is the count of the most prevalent class, O is the combined count of all other classes, and N is the total number of possible categories within the bias type (e.g., Gender: 2, Race: 5).

By dividing the score by N, the Dominance Ratio becomes easier to interpret, where a score approaching 0 indicates a more balanced distribution. This approach also eliminates the need for a predefined balance line, as the ratio inherently adjusts based on the number of categories. By combining Chi-Square tests and Normalized Dominance Ratios, we provide both statistical validation and intuitive visualization of bias within the UCF-101 dataset. This dual-metric approach allows us to not only detect biases but also understand their magnitude and impact.

3.3 Dataset Balancing

To mitigate the under-representation of certain demographic combinations in the UCF-101 dataset, we adopt a systematic approach to balance the dataset by adding new samples. Our strategy is to augment the dataset with carefully selected additional videos that counteract the observed imbalances.

Specifically, we employ an n-level factorial design to analyze the distribution of demographic attributes across various intersections. We begin with a one-way analysis (e.g., considering only race) and progressively extend our analysis to two-way (e.g., gender $\times$ age), three-way, four-way, and finally a full five-way analysis that simultaneously examines all five factors: gender, age, body shape, hair color, and race. For each level of analysis, the complete set of possible combinations is generated based on the factorial design, and these combinations are then compared against the observed distribution in the dataset. This comparison allows us to precisely identify which specific intersections are underrepresented and to quantify the number of additional samples required to achieve a balanced representation.

Once the underrepresented combinations are identified, we augment the dataset by scraping additional video data from YouTube. The additional videos are selected on the basis of their potential to counteract current biases and are integrated into the data set. After integration, we rerun our statistical tests to confirm that the demographic distribution is more balanced, thereby ensuring that the action recognition model is trained on a more fair and representative dataset.

3.3.1 Video Query Generation

To generalize our methodology, we approach data collection systematically. The high-level steps taken start from query generation which leads into gathering new data based on this queries from YouTube and ultimately filtering, downloading and assessing the new data. With the goal of being able to plug in any video dataset, it was necessary to start the query generation process from a dynamic standpoint, instead of statically writing queries based on a set of criteria. We do this by generating queries as a combination of variables from a predefined list of biases we aim to mitigate.

To quickly reiterate the bias categories we are focusing on, we detect and mitigate gender, race, age, hair color and body type. Our first iteration of the query generation step, we naturally view these bias categories as variables lists that we systematically iterate through and concatenate to produce a targeted, specific search for a specific bias we want to include. We take a targeted approach to search for niche, specific biases we aim to mitigate. Our query template consists of the following: "{bias category} + {action}". An example query resulting from this template is “woman PlayingPiano” or "white person JumpingJack". We use this template as our initial list of YouTube searches which result in 1717 separate searches.

3.3.2 Video Searching

For this research, we use YouTube as our go to platform for all new potential video data to be added to the debiased form of UCF-101. With the UCF-101 dataset, it is important for us to find the largest source in quantity of videos that are widely available so we can then be selective and filter down to potential new data. We do this by writing a simple but effective script that calls YouTubes own API to return the top 50 video results per search totaling $85,850$ videos for potential new data.

YouTube allows for up to 100 API calls per API key per day. To circumvent the API limits, we generate many API keys and have the script cycle through the keys after depletion detection. This enables us to have ample search opportunities for a variety of query templates that help us test out and refine relevant searches.

3.3.3 Potential Video Filtering

During the initial review stage of the potential videos the query results produced, we noticed large portions of data being either completely irrelevant to the search query or containing some form of NSFW content that we ended up discarding. In order to address this, we rework our YouTube query script to add additional parameters for filtering at the time of search.

YouTube's API offers a handful of filtering parameters to be applied to a given search call. For our study, we select to apply three filters: max duration, minimum views and safe search. We set the max duration to five minutes, minimum views to 100 and safe search to *moderate*.

The methodology behind setting the max duration to just five minutes pertains the compute expenses of editing long form videos into small clips. Videos taken from the UCF-101 dataset average about seven seconds per clip, allowing long form videos into our potential list only caused for waste and redundancy. The minimum views filter is set to filter out duplicate and irrelevant search results. With YouTube being so widely available to the public, we find many instances of re-uploaded videos with slightly different titles or videos where the titles simply do not match the contents of the videos. Applying the minimum view count is a way for us to naturally filter these irrelevant and duplicate uploads as these type of uploads do not have a high success rate on YouTube. Finally the safe search is set to moderate to filter out the potential NSFW content. After setting and testing these filters on time of search, our original $85,850$ results are filtered down to slightly over $14,000$ unique videos.

3.3.4 Quality Control

Each video form the search result is then attempted to be downloaded. Some videos are protected from being downloaded due to YouTube's policy and some videos are unavailable. Most of the videos in the original dataset are only a few seconds long. Multiple clips come from the same video, just with slightly different perspectives or lighting. To replicate this, the newly obtained videos are cut up into segments based on their scenes. To do this we use the adaptive detector from the pySceneDetect library. The adaptive detector performs rolling average on differences in HSV color-space. This change in scene will show the action but from a different perspective increasing the number of videos while avoiding duplication and maintaining quality. We then remove any instances where scene detection/splitting failed indicating the video is too noisy to clearly be split into scenes.

Of course, noise in the videos is expected, some scenes might be an intro, in-video advertisement, or some action that is not the desired action. To check for these type of situations we use the InternVideo2 model [19] and run inference. The model performs two sequential inferences: (1) it first detects whether a real human is present in the scene to minimize potential hallucinations in later steps, and (2) it then determines whether the specified action is actually being performed in the video. This is of paramount importance because at many time steps throughout these videos, the action is not happening. For the purposes of speed, after four positive scenes are detected, we move on to the next video.

To ensure the integrity of the dataset, we manually verify that in each of the selected clips, the action is happening and the model is not hallucinating. During this stage, we also trim scenes to ensure they are between 2 and 7 seconds in length. This also helps remove irrelevant parts of the scene, minimize the file sizes, and keep the video length consistent with the original dataset.

Finally, we run the videos through CLIP [14] or LLaVA [12] which perform five-shot video classification for three and two of the bias categories, respectively. CLIP was used for Age, Hair Color, and

Race, while LLaVA was used for Gender and Body Type. Each model was chosen for a particular category based on it's performance on the original dataset. The final labels were determined by a majority voting system. If there is a tie vote the label is marked as uncertain. Videos that have 3 out of 5 bias categories that were non-dominant in it's action, were kept. Any less were deemed to hurt the bias of the data set and were removed.

4 Experiments and Results

As previously mentioned in Section 3.1, we use both CLIP and LLaVA-OneVision to probe for biases within the dataset and the additional new found data. There were circumstances where one model performed better than the other in our particular study. CLIP performed much better in regards to Age, Hair color and Race, while LLaVA performed more consistent with Body Type and Gender. We found CLIP to be more reliable when the bias types contained more nuanced biases while LLaVA performed better with binary bias categories like body type and gender. For the following experiment and results, we kept the models compartmentalized and consistent with our initial testing when gathering the original biases before adding additional data. This is done to make sure we can maintain a fair comparison of the biases before and after the addition of the new data to cross reference if the bias was mitigated after our experiments.

4.1 Experiment

We designed a total of 170 search queries covering 10 distinct activity categories: Apply lip stick, Bench press, Diving, Drumming, Front crawl, Head massage, Hula hoop, Jumping jack, Playing guitar, and Playing piano. We ran these videos through our pipeline to test how effective our method is. From the collected data, 1530 raw videos were initially processed. After applying InternVideo2-based inference and manual verification, 531 clips were finalized. Among these, 373 clips satisfied our 3-out-of-5 rule, indicating strong agreement on the presence of the underrepresented groups.

4.2 Quantitative Results

Quantitative analysis is performed to evaluate the bias in the dataset before and after balancing. We successfully mitigated bias across ten unique action categories using our pipeline. We determine a successful bias mitigation if the added clip reduces the Dominance Ratio in at least 3 of the 5 bias categories we are working with. Below are the graph comparisons of the successful Dominance Ratio mitigation of the combined ten action categories before and after our clips were added.

It is inevitable some biases will increase while others decrease. It would be impossible to successfully add new clips that would only work to mitigate every bias category we aim to target. With the addition of our new clips into the dataset, the majority of the bias categories were mitigated and contain less bias overall across the action categories.

Table 1 and 2(chisq and dr scores) summarizes the key statistical metrics. Our results demonstrate a substantial reduction in bias, with improvements observed across the majority of demographic factors.

Demographic	χ^2 Value before	χ^2 Value after	Change
Gender	60.49	67.25	-6.76
Race	392.82	532.23	-139.41
Age	188.69	165.46	23.23
Hair Color	366.87	409.63	-42.76
Body	309.26	103.88	205.38

Table 1: Quantitative results showing χ^2 value of bias before and after balancing.

5 Conclusion

In this study, we introduced a structured pipeline for identifying and mitigating demographic bias in video datasets, using UCF-101 as a representative case. Our approach combined selective frame

Demographic	DR before	DR after	Change
Gender	1.48	1.45	0.03
Race	1.33	1.39	−0.06
Age	1.20	1.05	0.15
Hair Color	1.21	1.45	−0.24
Body	2.54	1.59	0.95

Table 2: Quantitative results showing Dominance Ratio (DR) of bias before and after balancing.

sampling, demographic annotation through CLIP and LLaVA, statistical bias measurement, and targeted augmentation via curated YouTube videos.

We applied this method to a subset of 10 out of the 101 action categories in UCF-101. While not all bias types improved uniformly, our results demonstrate an overall reduction in demographic bias across the selected categories. This validates the effectiveness of our method and highlights the importance of targeted data augmentation grounded in quantitative analysis.

Although time and computational constraints limited the scope of our study, the results suggest that scaling this pipeline to the full dataset could lead to significantly more balanced benchmarks. Our work lays the foundation for future efforts in dataset-level debiasing, aiming to enhance fairness and generalization in action recognition models.

References

[1] Yujia Bao and Regina Barzilay. Learning to split for automatic bias detection. *arXiv preprint arXiv:2204.13749*, 2022.

[2] Nitesh V Chawla, Kevin W Bowyer, Lawrence O Hall, and W Philip Kegelmeyer. Smote: synthetic minority over-sampling technique. *Journal of artificial intelligence research*, 16: 321–357, 2002.

[3] Georgios Douzas, Fernando Bacao, and Felix Last. Improving imbalanced learning through a heuristic oversampling method based on k-means and smote. *Information sciences*, 465:1–20, 2018.

[4] Justin Engelmann and Stefan Lessmann. Conditional wasserstein gan-based oversampling of tabular data for imbalanced learning. *arXiv preprint arXiv:2008.09202*, 2020.

[5] Haibo He, Yang Bai, Edwardo A Garcia, and Shutao Li. Adasyn: Adaptive synthetic sampling approach for imbalanced learning. In *2008 IEEE international joint conference on neural networks (IEEE world congress on computational intelligence)*, pages 1322–1328. Ieee, 2008.

[6] Phillip Howard, Avinash Madasu, Tiep Le, Gustavo Lujan Moreno, Anahita Bhiwandiwalla, and Vasudev Lal. Socialcounterfactuals: Probing and mitigating intersectional social biases in vision-language models with counterfactual examples. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 11975–11985, 2024.

[7] Kai Hu, Feng Gao, Xiaohan Nie, Peng Zhou, Son Tran, Tal Neiman, Lingyun Wang, Mubarak Shah, Raffay Hamid, Bing Yin, et al. M-llm based video frame selection for efficient video understanding. *arXiv preprint arXiv:2502.19680*, 2025.

[8] Nikita Jaipuria, Xianling Zhang, Rohan Bhasin, Mayar Arafa, Punarjay Chakravarty, Shubham Shrivastava, Sagar Manglani, and Vidya N Murali. Deflating dataset bias using synthetic data augmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 772–773, 2020.

[9] Vishwa Karia, Wenhao Zhang, Arash Naeim, and Ramin Ramezani. Gensample: A genetic algorithm for oversampling in imbalanced datasets. *arXiv preprint arXiv:1910.10806*, 2019.

[10] Kimmo Karkkainen and Jungseock Joo. Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *Proceedings of the IEEE/CVF winter conference on applications of computer vision*, pages 1548–1558, 2021.

[11] Ludmila I. Kuncheva. *Combining Pattern Classifiers: Methods and Algorithms*. Wiley-Interscience, 2004.

[12] Bo Li, Yuanhan Zhang, Dong Guo, Renrui Zhang, Feng Li, Hao Zhang, Kaichen Zhang, Peiyuan Zhang, Yanwei Li, Ziwei Liu, et al. Llava-onevision: Easy visual task transfer. *arXiv preprint arXiv:2408.03326*, 2024.

[13] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *ACM computing surveys (CSUR)*, 54(6):1–35, 2021.

[14] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. URL `https://arxiv.org/abs/2103.00020`.

[15] Seonguk Seo, Joon-Young Lee, and Bohyung Han. Unsupervised learning of debiased representations with pseudo-attributes. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 16742–16751, 2022.

[16] Khurram Soomro, Amir Roshan Zamir, and Mubarak Shah. Ucf101: A dataset of 101 human actions classes from videos in the wild. *arXiv preprint arXiv:1212.0402*, 2012.

[17] Antonio Torralba and Alexei A Efros. Unbiased look at dataset bias. In *CVPR 2011*, pages 1521–1528. IEEE, 2011.

[18] Angelina Wang, Alexander Liu, Ryan Zhang, Anat Kleiman, Leslie Kim, Dora Zhao, Iroha Shirai, Arvind Narayanan, and Olga Russakovsky. Revise: A tool for measuring and mitigating bias in visual datasets. *International Journal of Computer Vision*, 130(7):1790–1810, 2022.

[19] Yi Wang, Kunchang Li, Xinhao Li, Jiashuo Yu, Yinan He, Chenting Wang, Guo Chen, Baoqi Pei, Ziang Yan, Rongkun Zheng, Jilan Xu, Zun Wang, Yansong Shi, Tianxiang Jiang, Songze Li, Hongjie Zhang, Yifei Huang, Yu Qiao, Yali Wang, and Limin Wang. Internvideo2: Scaling foundation models for multimodal video understanding, 2024. URL `https://arxiv.org/abs/2403.15377`.

[20] Wenyu Zhang, Qingmu Liu, Felix Ong Wei Cong, Mohamed Ragab, and Chuan-Sheng Foo. Universal semi-supervised domain adaptation by mitigating common-class bias. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 23912–23921, 2024.

[21] Zaiying Zhao, Soichiro Kumano, and Toshihiko Yamasaki. Language-guided detection and mitigation of unknown dataset bias. *arXiv preprint arXiv:2406.02889*, 2024.

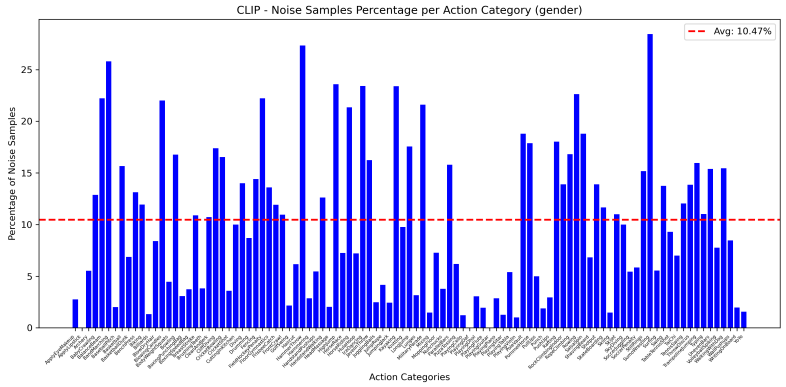

Figure 3: CLIP gender bias noise percentage

A Supplementary Material

In the following section, we present additional material regarding our statistical findings in respect to the overall UCF-101 Action Dataset as well as the biases found within. Specifically, we display and discuss some statistical breakdown of the overall UCF-101 dataset as well as the results of our CLIP and LLaVA experiments. We also provide further analysis on the quantitative metrics of the biases discovered.

As discussed in 3.1, we implement a smoothing method for both the CLIP and LLaVA bias detections to ensure consistent and denoised results. Below are the noise sample percentage per action category per bias group. A notable portion of the bias detections contained some sort of noise which was denoised through our process. We show these graphs below to display the importance of this denoise step for each of the bais categories for both CLIP and LLaVA detections. We used the same fixed seed for the frame selections for both models when prompting the models for the biases within. This allows for a fair comparison in regards to the noise detections, dominance ratio and overall bias detections.

A.1 CLIP Results

We first start off by displaying the noise percentage of the CLIP results for both the Gender and Race bias detections. We display just these two to give a better understanding of the noise within the bias detection step relative to each model. The noise percentages were similar across all bias categories relative to the model, so these two give a good understanding of the baseline metrics.

In figures 3 and 4, there is a relatively large portion of noise samples prevelant. We see an average of 10.47% and 23.57% of the samples containing noise in regards to Gender and Race respectively. There was a significant amount of noise within the CLIP bias detections. With our methodology of random frame sampling, there is a lot of room for noise which can be very dependent on the randomly sampled frame. Noise can be a product of frame lighting, camera angle, frame clarity and many other factors.

CLIP Dominance Ratio Results We next go into the Dominance Ratio findings for the CLIP model bias detection. The figures 5, 6, 7, 8, 9 relate to the dominance ratio findings and were taken after the data was smoothed and denoised.

A.2 LLaVA Results

We again start off the LLaVA results by displaying the noise percentage for both the Gender and Race bias detections in respect to the LLaVA model. We again display just these two to give a better understanding of the noise within the bias categories and these two give a good understanding for a baseline metric.

Across all bias categories, the LLaVA model produced more consistent, less noisy results. The percentage of noise per action category for Gender 10 and Race 11 detections were 8.54% and

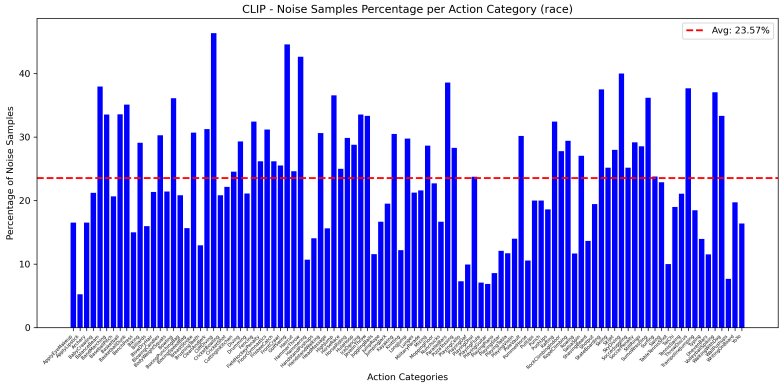

Figure 4: CLIP race bias noise percentage

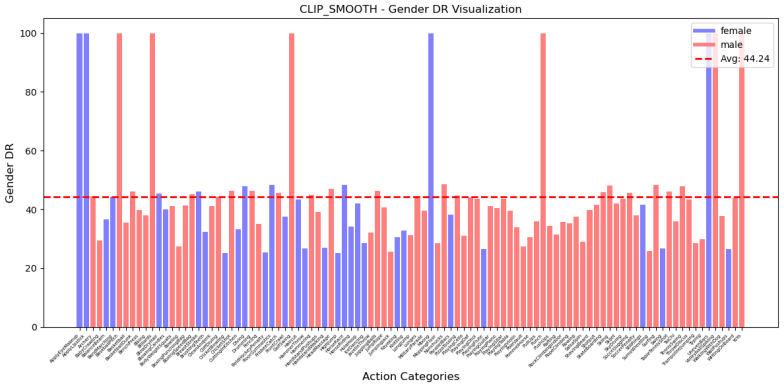

Figure 5: CLIP Gender Dominance Ratio

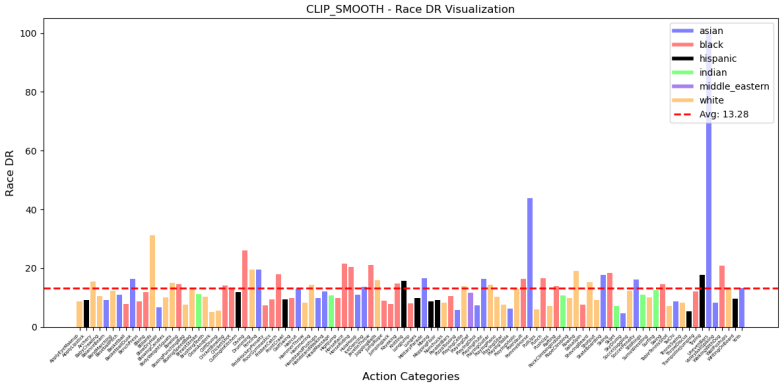

Figure 6: CLIP Race Dominance Ratio

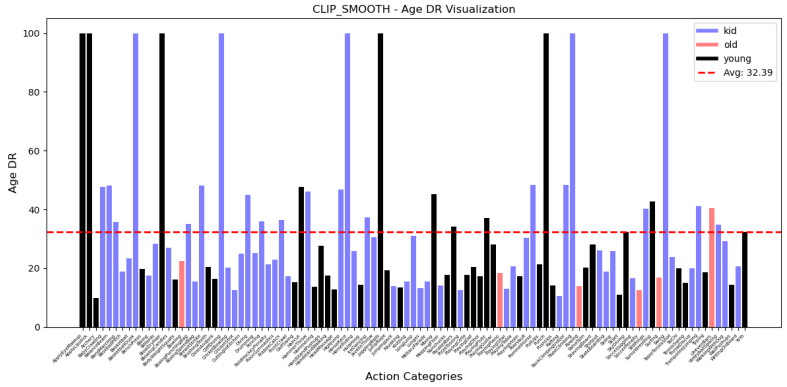

Figure 7: CLIP Age Dominance Ratio

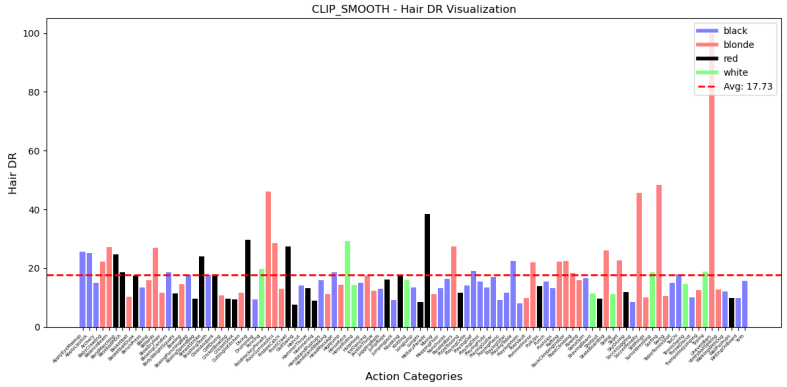

Figure 8: CLIP Hair Dominance Ratio

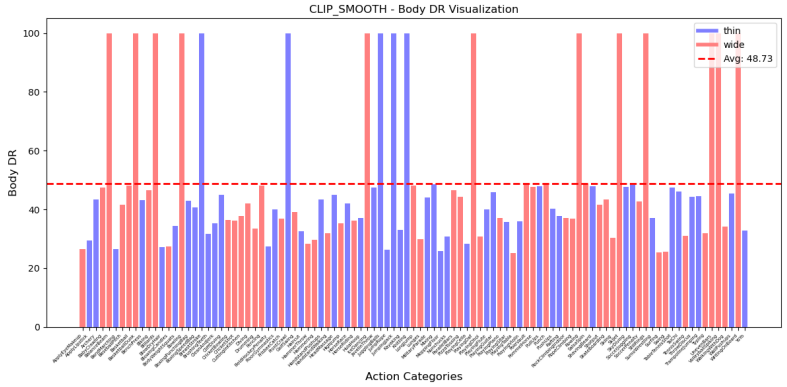

Figure 9: CLIP Body Dominance Ratio

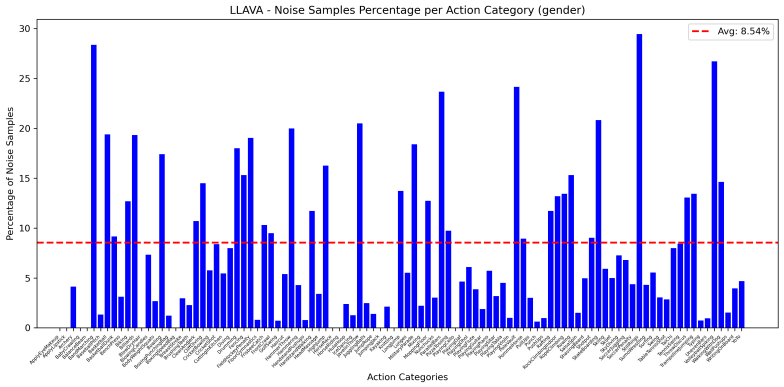

Figure 10: LLaVA Gender Noise Percentage

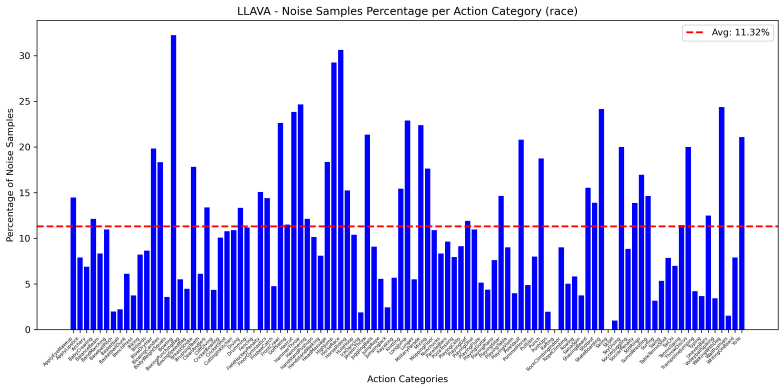

Figure 11: LLaVA Race Noise Percentage

11.32% respectively. It is worth noting that the LLaVA model detections contained less noise than the CLIP detections, although still containing some. Regardless of the model used, it was necessary for us to detect and address the noise during the statistical analysis portion.

LLaVA Dominance Ratio Results We next go into the Dominance Ratio findings, this time for the LLaVA model bias detection. The following dominance ratio figures 12, 13, 14, 15, 16 were taken again after the data was smoothed and denoised.

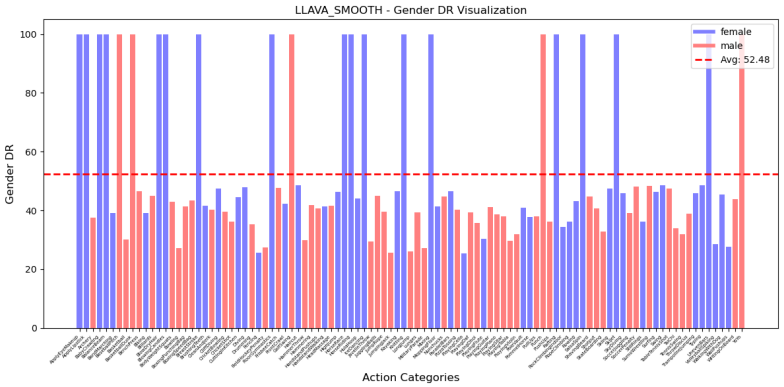

Figure 12: LLaVa Gender Dominance Ratio

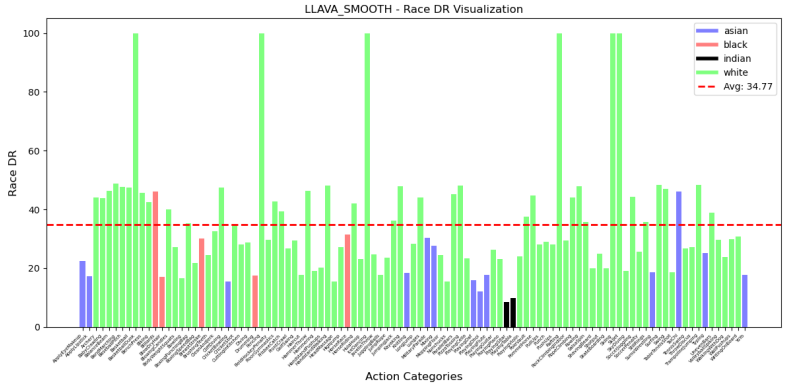

Figure 13: LLaVa Race Dominance Ratio

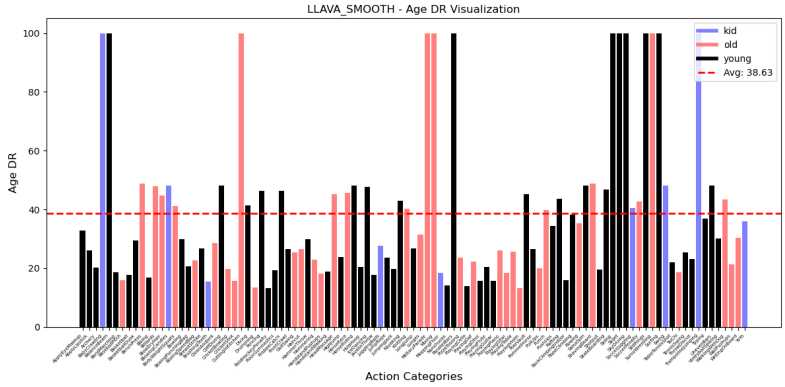

Figure 14: LLaVa Age Dominance Ratio

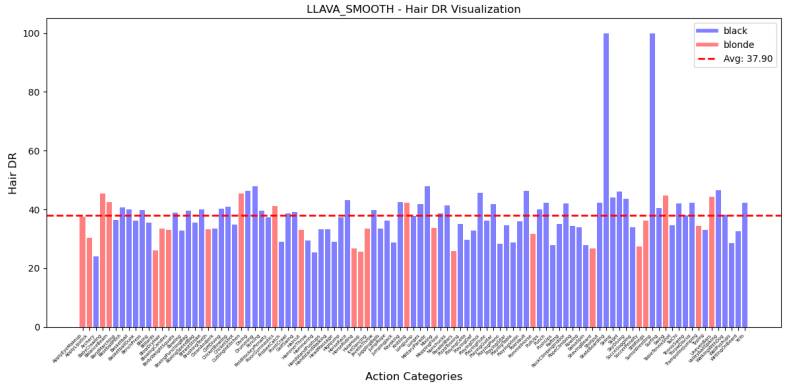

Figure 15: LLaVa Hair Dominance Ratio

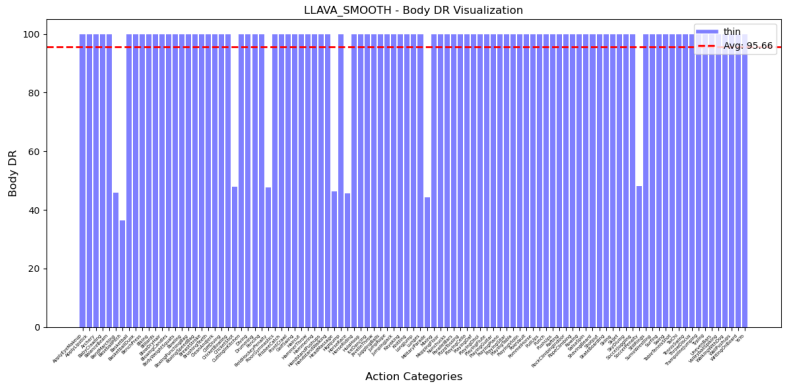

Figure 16: LLaVa Body Dominance Ratio

A.3 Dominance Ratio Results

The following figures are the dominance ratio and chi score graphs for some of the successfully mitigated bias categories, comparing the dominance before our added clips and after our added clips.

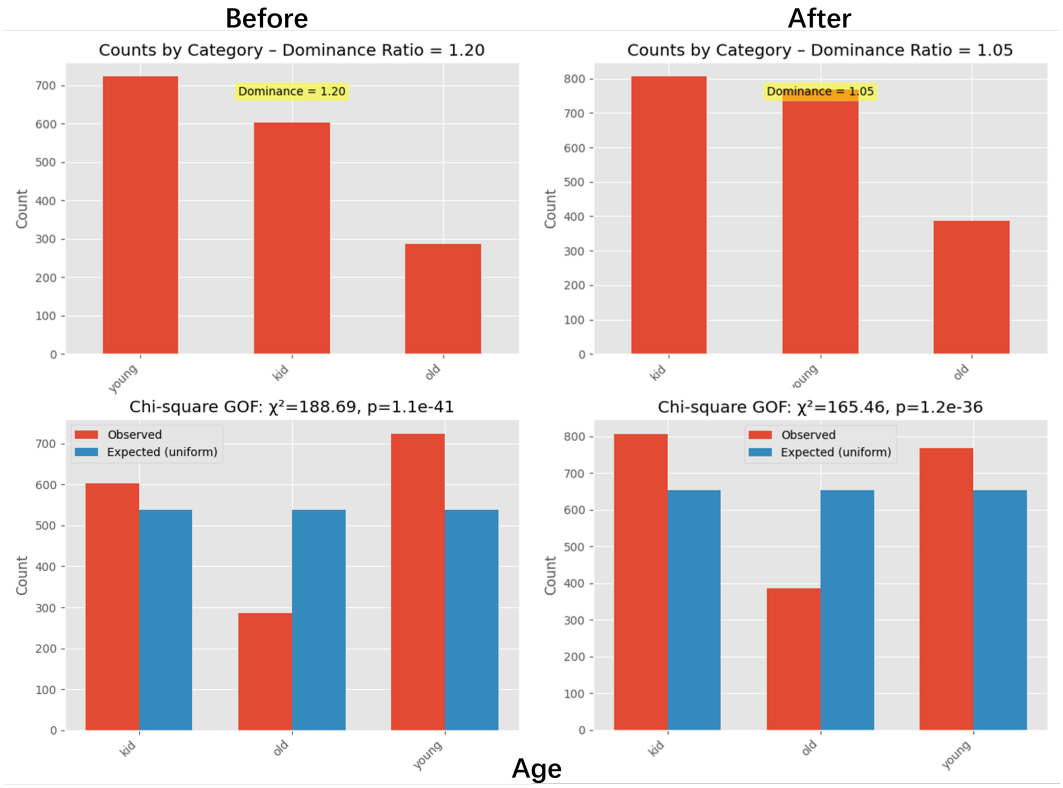

Figure 17: Dominance ratio of Age before (left) vs after (right)

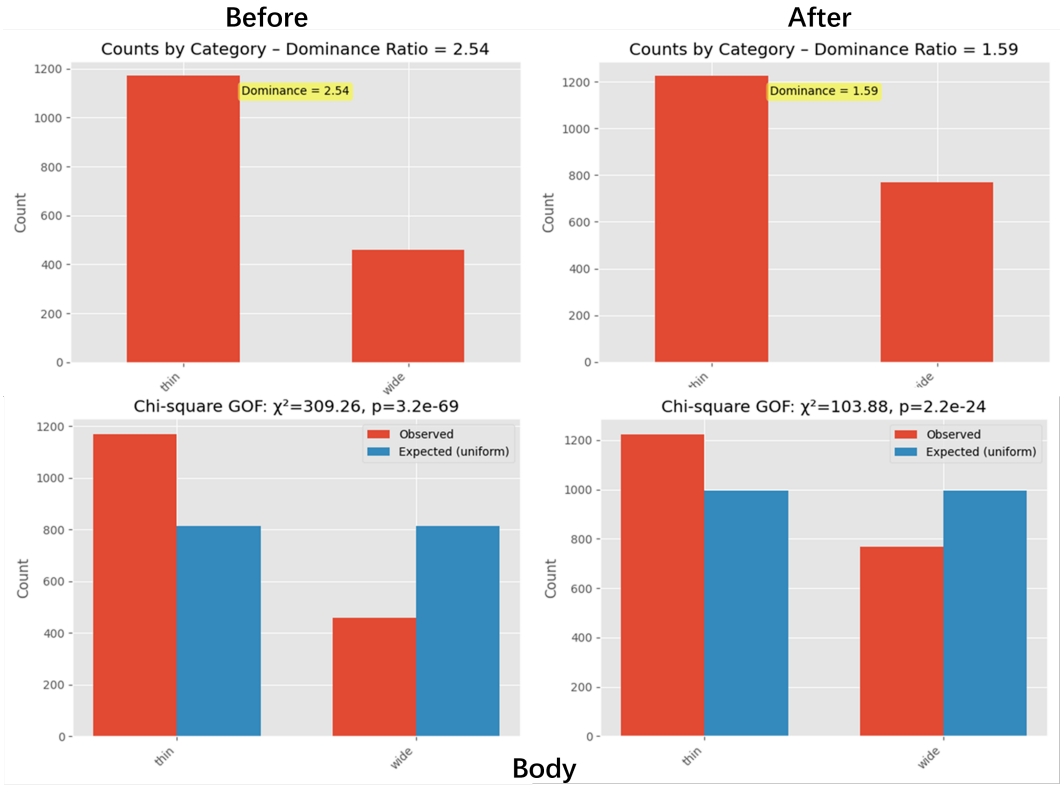

Figure 18: fig:Dominance ratio of Body before (left) vs after (right)

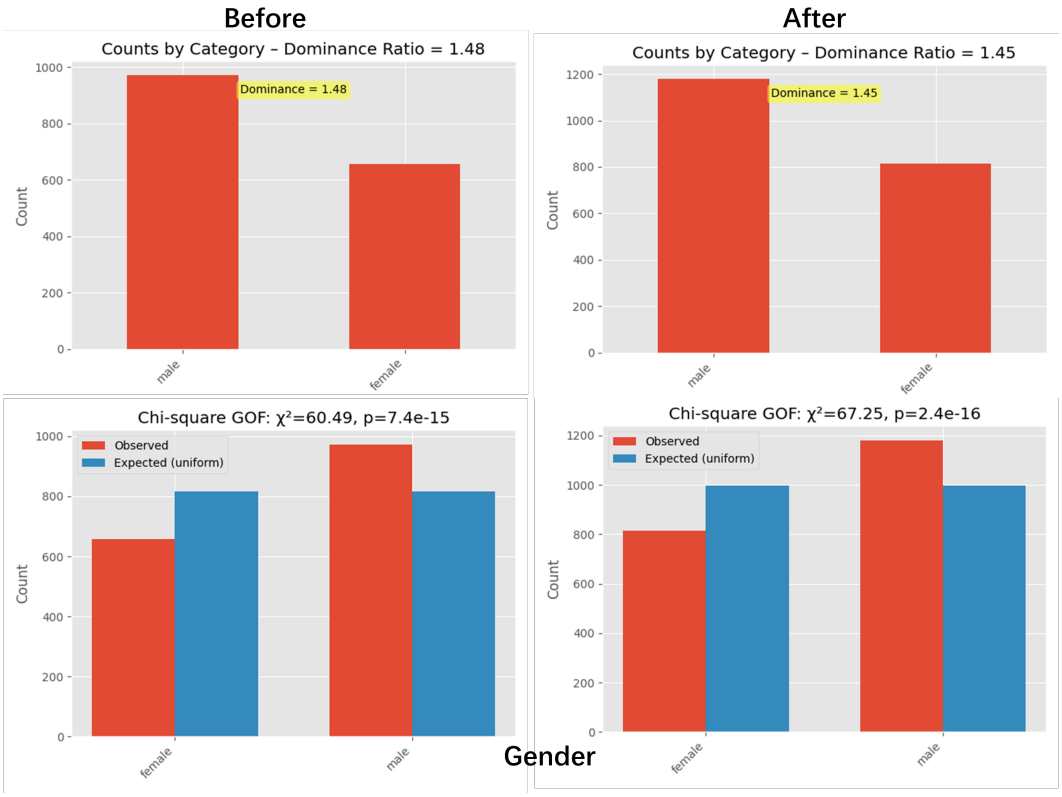

Figure 19: Dominance ratio of Gender before (left) vs after (right)

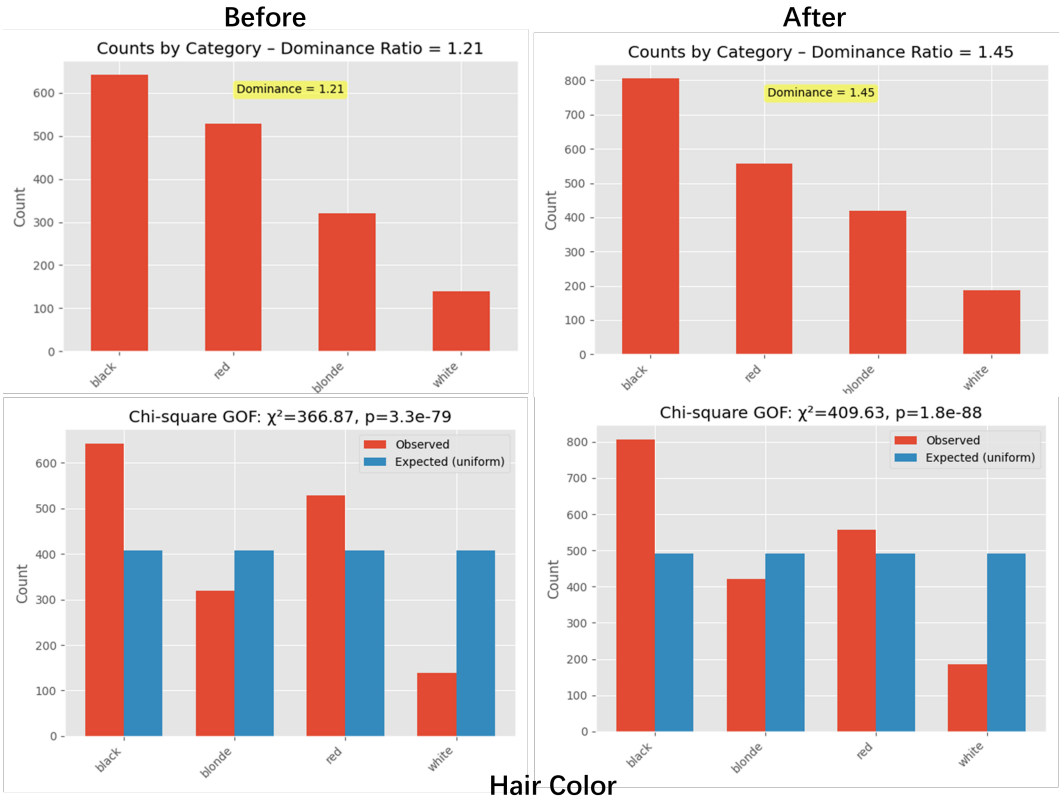

Figure 20: Dominance ratio of Hair Color type before (left) vs after (right)

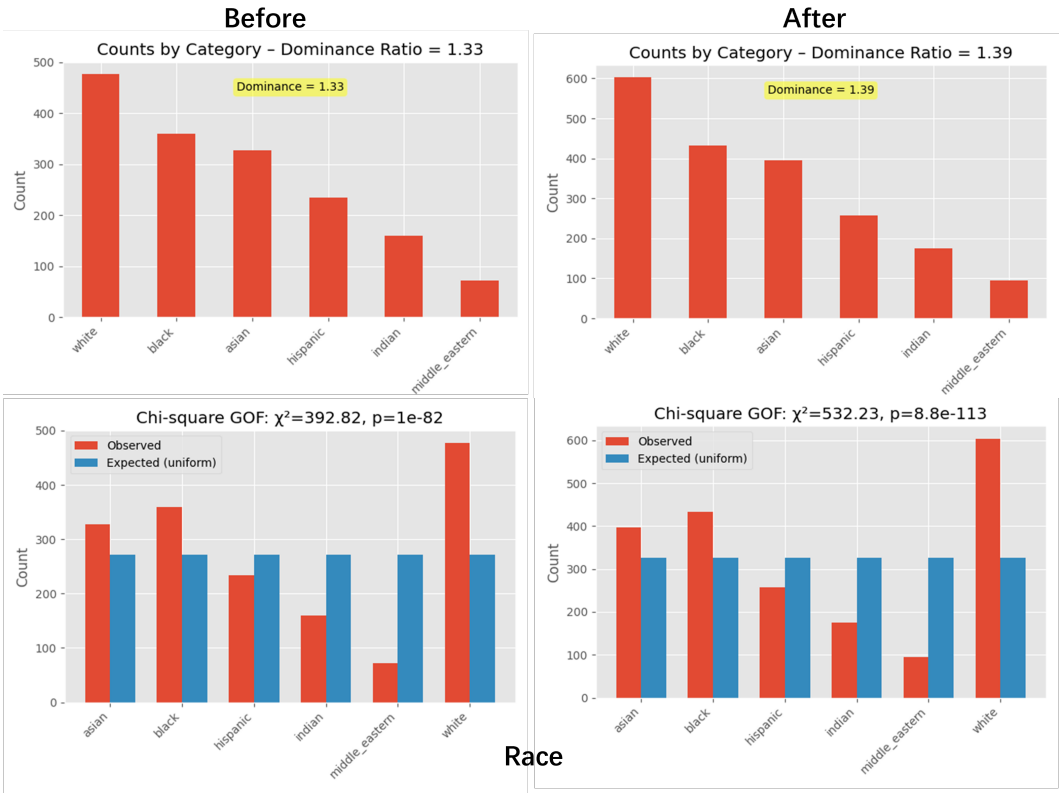

Figure 21: Dominance ratio of Race type before (left) vs after (right)