

D3G: Diverse Demographic Data Generation Increases Zero-Shot Image Classification Accuracy within Multimodal Models

Javon Hickmon

Department of Computer Science, University of Washington, Seattle WA
javonh@cs.washington.edu

Abstract

Image classification is a task essential for machine perception to achieve human-level image understanding. Multimodal models such as CLIP, have been able to perform well on this task by learning semantic similarities across Vision and Language; however, despite these advances, image classification is still challenging task. Models with low capacity often suffer from underfitting and thus underperform on fine-grained image classification. Along with this, it is important to ensure high-quality data with rich cross-modal representations of each class, which is often difficult to generate. When datasets do not enforce balanced demographics, the predictions will bias toward the more represented class, while others will be neglected. We focus on how these issues can lead to harmful bias for zero-shot image classification, and explore how to combat these issues in demographic bias. We propose Diverse Demographic Data Generation (D3G), a training-free, zero-shot method of boosting classification accuracy while reducing demographic bias in pre-trained multimodal models. With this method, we utilize CLIP as our base multimodal model, and Stable Diffusion XL as our generative model. We demonstrate providing diverse demographic data at inference time improves performance for these models, and explore the impact of individual demographics on the resulting accuracy metric.

1 Introduction

Deep Learning systems have been a promising new paradigm in the field of image classification. Vision-Language systems are able to utilize multiple modalities to create models that are more generalizable on a broad range of downstream tasks. Despite this performance, issues such as data redundancy, noise, and class imbalance are just a few of the many difficulties that arise from collecting large amounts of training data. Multimodal models in particular, require large amounts of high-quality training data with rich cross-modal representations in order to perform well compared to their unimodal counterparts. These models leverage the massive amounts of image text pairs available online by learning to associate the images with their correct caption, leading to greater flexibility during inference (Pratt et al. 2023); however, class imbalances can often lead to gender and racial bias depending on the desired task.

Our code is available at: <https://github.com/jjhickmon/D3G>

Existing public face image datasets are strongly biased toward Caucasian faces, meanwhile other races (i.e., Latino) are significantly underrepresented Figure 4. As a result, the models training from such datasets suffer from inconsistent classification accuracies, limiting the applicability of such systems work predominantly on White racial groups Karkainen and Joo (2021). This means that minority sub-populations can potentially be further marginalized when applied to certain downstream tasks without calibration. This is a core tenet of machine learning: biased data produces biased models.



Figure 1: Images Generated with D3G for Race 4

It is important to note that we acknowledge that not all bias is harmful and often is necessary for models to generalize. This is why within this work, we focus on demographic bias which can frequently have harmful implications when applied within society. Image classification is particularly pressing, because it is the core of a myriad of computer vision tasks. Facial recognition, object detection, image search, content moderation, sentiment analysis, and many more tasks are grounded in accurate image classification

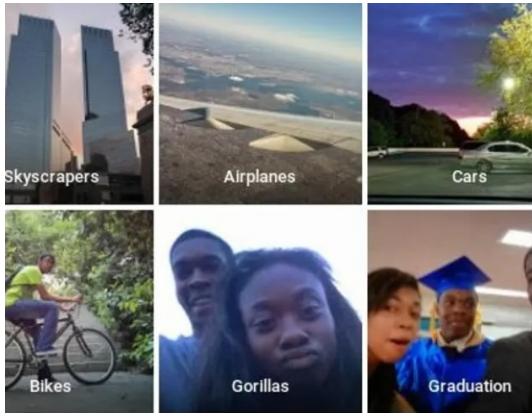


Figure 2: The image of Google Photos misclassifying Black people within the Photos application.

systems. This is compounded by the fact that many widely used Foundational Models require multiple modalities, such as DALL-E [Ramesh et al. \(2021\)](#) and Stable Diffusion [Rombach et al. \(2022\)](#). In order to train these models, other models such as CLIP [Radford et al. \(2021\)](#) are used to classify the data that the model will be trained on, in order to enforce a strong cross-modal correlation. This means that demographic biases will be compounded as the images continue to be utilized in training processes.

When there are strong harmful demographic biases, these models can cause tremendous harms. One example was when Google Photo's classified Black people as gorillas within a user's album as shown in Figure 2. This is the direct result of demographic bias. The dataset used to quantify the classification accuracies of Google's model, likely contained the standard biases where images of non-Caucasian faces are underrepresented Figure 4. This leads to a misleading evaluation, which was likely why this model was deployed to the public with this significant issue.

In this work, we explore how demographic bias affects image classification accuracies for multimodal models. We also propose D3G, a zero-shot, training-free framework to balance demographic bias and boost classification accuracies for multimodal models used for image classification.

2 Related Work

2.1 Model Ensembling for Image Classification

Many state-of-the-art techniques for image classification leverage a methodology known as model ensembling. Ensemble learning broadly is an approach that aims to improve performance by combining the predictions of multiple models. There are many such ensemble learning methods, but the one most relevant to our proposed technique is called bagging.

Bagging predictors originally published by [Breiman](#)

(1996) introduced "bootstrap aggregating" or bagging, the ensemble learning method that combines multiple models trained on different subsets of the training data. The formulation is presented as follows: A learning set of L consists of data $(y_n, \mathbf{x}_n), n = 1, \dots, N$ where the y 's are class labels, assume we form a predictor $\phi(\mathbf{x}, L)$ where if the input is \mathbf{x} , we predict y by $\phi(\mathbf{x})$. Now suppose we form a sequence of replicate learning sets $L^{(B)}$ each consisting of N observations drawn at random with replacement from L . Since y is a class label in our scenario, the predictions from each of the predictors $\phi(\mathbf{x}, L^{(B)})$ will vote to form $\phi_B(\mathbf{x})$, the aggregated final prediction. Bagging both empirically and theoretically prove improves accuracy for a given set of weak classifiers or "weak learners." This technique effectively replicates our proposed method within a controlled environment. In the implementation of D3G, we are employing a strategy similar to bagging, but across modalities and with generated data. Our goal is based on the theoretical guarantees of bagging, where even though each model is trained on a subset of the data, the aggregation of the predictions begins to approximate the true distribution of the data.

A model that applies this concept of model ensembling was introduced in [Learning to Navigate for Fine-grained Classification](#) by [Yang et al. \(2018\)](#). This is a state-of-the-art paper that attempts to reduce misclassification rates by developing a model called NTS-Net (Navigator-Teacher-Scrutinizer Network) to teach itself methods of identifying and scrutinizing fine-grained image details. The Navigator navigates the model to focus on the most informative regions (denoted by yellow rectangles in Figure 3), while the Teacher evaluates the regions proposed by the Navigator and provides feedback. After that, the Scrutinizer scrutinizes those regions to make predictions. NTS-Net achieves high classification accuracy on a pre-defined dataset; however, we wanted to explore if these foundational concepts could be applied to a training-free, zero-shot environment. In order to achieve this, we leverage pretrained open-vocabulary models with advanced attention mechanisms to discriminate the fine-grained features of a given image, in order high performance on a variety of classes without additional training.

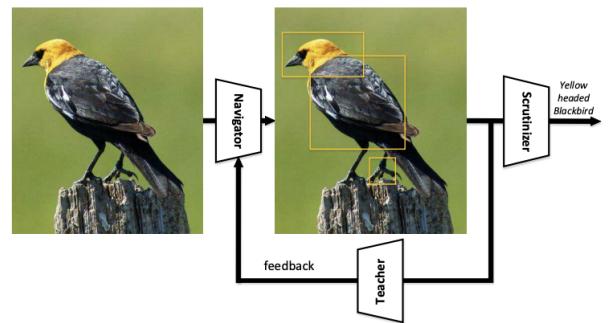


Figure 3: The NTSNet Architecture ([Yang et al. 2018](#))

Similarly, within the paper *Sus-X: Training-Free Name-Only Transfer of Vision-Language Models* by [Udandarao, Gupta, and Albanie \(2022\)](#) achieves state-of-the-art zero-shot classification results on 19 benchmark datasets, outperforming other training-free adaptation methods. It also demonstrates strong performance in the training-free few-shot setting, surpassing previous state-of-the-art methods. This paper is focused on general image classification improvements; however we aim to explore how this idea of synthetic support set generation affects the fairness of predictions from a classification model. We will employ a similar strategy but also explore how to offset existing harmful biases within the zero-shot setting.

2.2 Data Filtering and Generation

Neural Priming for Sample-Efficient Adaptation by [Wallingford et al. \(2024\)](#) proposes a technique to adapt large pretrained models to distribution shifts. This paper demonstrates that we can leverage an open-vocabulary model’s own pretraining data in order to improve performance on downstream tasks. Even though we don’t aim to utilize the model’s training data in our method, the generated images will likely be sampled from a similar distribution as multimodal model. This paper shows that even if that is the case, we can still use filtering and guidance in order to improve performance. In our case, our custom prompting method plays the role of guiding the image generation process, resulting in the same empirical performance improvements.

DATACOMP: In search of the next generation of multimodal datasets by [Gadre et al. \(2024\)](#) has completely different goals from Neural Priming, but achieves them in a similar manner. The paper introduces DataComp, which is essential a test bed for dataset-related experiments that contains 12.8 billion image-text pairs retrieved from Common Crawl. Upon retrieving this pool, they proceed to train a new clip model with a fixed architecture and hyper-parameters. The paper concludes that CommonPool and LAION-2B are comparable with the same filtering. This means that image-based filtering and CLIP score filtering excels on most tasks, and can be effectively used to retrain other models. Despite this, the paper mentions that they found demographic biases in models trained using their pool, but their goal was not to reduce these harmful biases. In this paper we aim to offset this demographic bias found in models trained on large-scale filtered data pools such as DataComp.

2.3 Ethics and Fairness

The FairFace dataset and classifier were first published in *FairFace: Face Attribute Dataset for Balanced Race, Gender, and Age for Bias Measurement and Mitigation* by [Karkkainen and Joo \(2021\)](#). This project focused on creating a dataset and classifier that were balanced across race, gender and age as shown in Figure 4. This balance is crucial because the paper demonstrates that the balance allows for improved generalization classification performance on the defined demographics, even on novel datasets that contain

more non-White faces than typical datasets. The fact the simply balancing these demographics allows for increased accuracy and generalizability is extremely important. This is the core of D3G, and FairFace shows that balancing demographics results in performance improvements. The primary difference is that we aim to show similar improvements without any additional training. Alongside creating a balanced dataset, they also demonstrated their classifier produces balanced accuracy across the specified demographics, which is crucial because I use this classifier to create new labels for the IdenProf dataset.

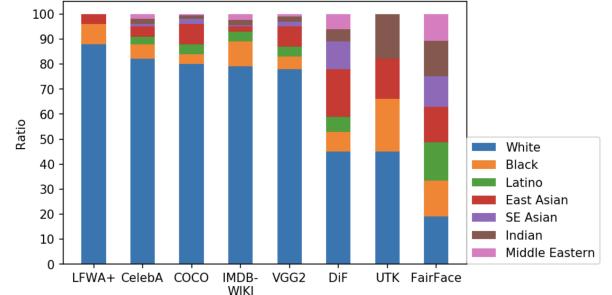


Figure 4: Racial compositions in face datasets ([Karkkainen and Joo 2021](#))

3 Methods

We aim to create an ensemble of models to improve multimodal image classification accuracy, especially for models that are trained on data with a class imbalance. We test this method on standard benchmark datasets, such as ImageNet as shown in Figure 6, then we expand our technique to classify demographic-focused datasets. CLIP (Contrastive Language-Image Pretraining) ([Radford et al. 2021](#)) will be used for image-to-text retrieval, and Stable Diffusion XL 1.0 ([Podell et al. 2023](#)) for image generation. Our approach is as follows:

3.1 Datasets

For all of the results shown in this paper, we classify images from the IdenProf test dataset. We selected this dataset because it provides a simple, applicable downstream task and because all of the images were collected and filtered by hand via Google Image search. Each image in the dataset can belong to one of ten classes: Chef, Doctor, Engineer, Farmer, Firefighter, Judge, Mechanic, Pilot, Police, or Waiter. In total there are 2,000 images for testing, with 200 images for each class. Finally, it is important to note the demographic distribution published by the dataset authors. The IdenProf dataset consists of 80.6% male subjects, and 19.4% female. Along with this, 91.1% of the people within the dataset are White, while 8.9% are of another race. The dataset author also noted that there were more images of Asian and White people obtainable, when compared to that of Black people. Similarly, there were more images of men

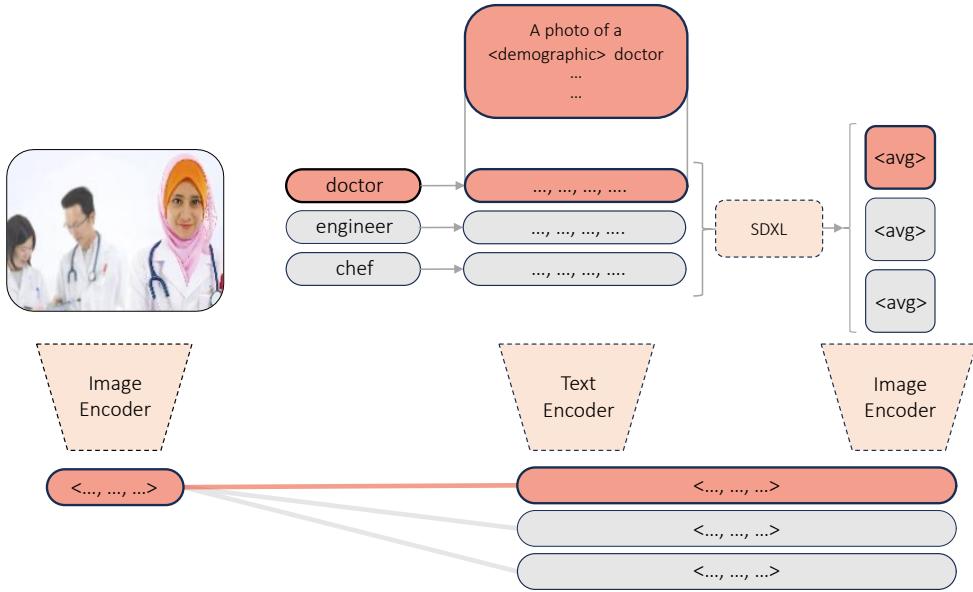


Figure 5: The D3G Framework

obtainable than of women. This reflects this demographic biases discussed previously.

Along with IdenProf, we also leverage information collected from the FairFace dataset [Karkainen and Joo \(2021\)](#). This dataset defines common demographics and forms them into classification categories. The authors constructed their dataset containing 108,501 images, and even though we do not utilize this dataset within this paper, the demographic information is still useful. We leverage the classification model that they trained on their own dataset. As a result, the results from the classifier are highly balanced and less likely to contain demographic bias. We use this classifier to assign additional labels to the images within IdenProf. There are 3 primary demographics that will be assigned as labels: race, gender, and age. Along with this, the race category has two versions, with race 4 being coarse-grained with only four races to choose from, and race 7 being fine-grained with 7 races to choose from. Combining the classes from IdenProf and FairFace, every image in the dataset can be classified with any of the labels identified in Table 1.

3.2 Creating Prompts

To generate our prompts, we leverage a set of templates constructed based on demographics identified in Table 1. These templates are designed to expose and leverage a specific demographic bias, based on the whatever image is currently being classified. For instance, if we were attempting to classify the profession of the person within the image, our prompts would be as shown in Table 2. This process is pictured within Figure 5.

Class	Values
profession	Chef, Doctor, Engineer, Farmer, Firefighter, Judge, Mechanic, Pilot, Police, or Waiter
race 7	White, Black, Indian, East Asian, Southeast Asian, Middle Eastern, and Latino
race 4	White, Black, Indian, Asian
gender	Male, Female
age	0-2, 3-9, 10-19, 20-29, 30-39, 40-49, 50-59, 60-69, 70+

Table 1: All potential classes for an image from IdenProf

3.3 Generate Class Images

Upon creating diverse demographic prompts from templates for each of the classes, each of these prompts are used to generate an image. We employ Stable Diffusion XL, a diffusion-based image generation model, to conditionally generate an image of each class in the dataset. Our result will be images that emphasizes the diverse demographics between the classes. For standard D3G we will generate 1 image per prompt then average the embeddings of all the prompts, and for average image D3G we will generate 5 images per prompts, then perform the same process of averaging the embeddings of these images. Generating these diverse images is crucial because our goal is to combat the issue of prediction bias by generating diverse images and utilizing them for the next step when predicting labels. This step is depicted by images on the right of Figure 6.

3.4 Weighted Sum

Using the prompts created earlier, we start the classification phase. We use the image and text encoders from CLIP ViT-

Demographic	Prompt	Text
Profession	"A photo of a <prof>"	A photo of a doctor
Race 7	"A photo of a <race> <prof>"	A photo of a white doctor
Race 4	"A photo of a <race> <prof>"	A photo of a white doctor
Gender	"A photo of a <gender> <prof>"	A photo of a male doctor
Age	"A photo of a <age> year old <prof>"	A photo of a 30-39 year old doctor

Table 2: Example diverse demographic texts for classifying *profession*. Note that each prompt starts with "A photo of a," and that all of the correct nouns and adjectives are added to the prompts as shown in the right column. More examples are provided in Section 10.

L/14, our multimodal model, in order to get the embeddings for the generated images. Upon getting these embeddings, we scan values from 0 to 1 using a step value of 0.01 in order to find an optimal weight to create a weighted sum of the text and image embeddings. The text embedding will have a weight of w while the image embedding is weighted by $1 - w$. This step allows us to bridge the semantic gap between text and images, because images are always closer in embedding space to one another than text. After performing this step, we will get a new embedding that represents the weighted combination of the text and image embeddings.

3.5 Classification

Finally, we get the embedding of the query image by passing it through the CLIP image encoder. At this point, we simply get the cosine similarity between the query image embedding, and the combined image-text embeddings from each class. In order to classify the image, we just get the highest similarity score and use that class as the prediction. This step is depicted by the blue arrows within Figure Figure 6.

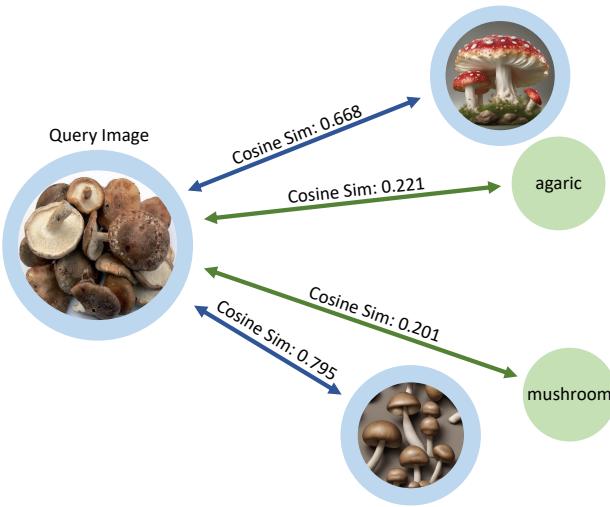


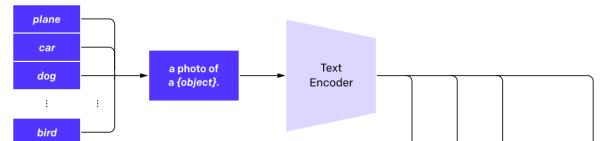
Figure 6: A demo example of D3G on difficult fine-grained classes from the ImageNet dataset (Note: we utilized ImageNet for this example to showcase the fine-grained classification capabilities of D3G. IdenProf does not have such fine-grained classes).

4 Results

4.1 Metrics

We choose to use top-1 accuracy as the standard metric for our results. We selected this metric for a variety of reasons, the most prominent being that this paper aims to increase zero-shot classification accuracy. There are many metrics that represent zero-shot accuracy; however, top-1 accuracy is the most common.

2. Create dataset classifier from label text



3. Use for zero-shot prediction

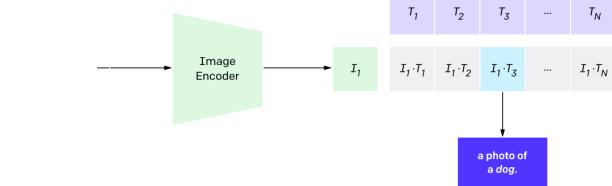


Figure 7: The process of classifying an image with CLIP at inference-time (Radford et al. 2021).

4.2 Evaluation Breakdown

We study three primary classification methods in this paper. CLIP ViT-L/14 is our baseline, the standard method of multimodal classification as outlined in Figure 7. The second method of image classification is Standard D3G as shown in Figure 5. In this method, for every class in our dataset, we generate one prompt for each of the specified demographics, then use these prompts to generate images and average their embeddings. Finally, our third method is Average Image D3G. This is the exact same process as Standard D3G; however, instead of generating one image per demographic prompt, we will generate 5, then average the embeddings of all the prompts for a given class.

Along with our three classification methods, we outline 5 prompting strategies when creating our demographic

Demographic	Method	Profession	Race 7	Race 4	Gender	Age
Profession	CLIP	95.14	94.73	95.22	96.52	94.81
	Standard D3G	95.54	95.22	95.30	96.52	95.06
	Average Image D3G	95.87	95.62	95.38	96.76	95.54
Race 7	CLIP	44.65	28.20	-	28.61	25.69
	Standard D3G	45.38	31.85	-	32.90	30.96
	Average Image D3G	45.46	32.33	-	33.55	32.25

Table 3: Results when classifying the specified demographic of the people within the IdenProf dataset. The far left column shows the demographic that will be classified. The second column on the left dictates the method used for classification, while the other columns dictate the prompting structure as discussed in Section 4.2. Additional results are shown in Section 10.

Demographic	Method	Profession	Race 7	Race 4	Gender	Age
Profession	Standard D3G	0.85 / 0.15	0.84 / 0.16	0.90 / 0.10	0.91 / 0.09	0.90 / 0.10
	Average Image D3G	0.71 / 0.29	0.74 / 0.26	0.84 / 0.16	0.91 / 0.09	0.67 / 0.33
Race 7	Standard D3G	0.90 / 0.10	0.68 / 0.32	-	0.68 / 0.32	0.67 / 0.33
	Average Image D3G	0.92 / 0.08	0.69 / 0.31	-	0.67 / 0.33	0.68 / 0.32

Table 4: The weight values used to achieve the results in Table 3. For each evaluation, the left value is the text embedding weight, and the right is the image embedding weight. CLIP is not included because no images are weighted with the text embeddings. Note that the sum of the text and image weights for a given evaluation should equal 1.

prompts within the D3G framework as shown in the top row of Table 3. It is important to note that for all of these prompting strategies, they *add* demographic information, in addition to the specified classification category. For instance, as shown in Table 2, when the task is to classify profession, we can add information regarding race or gender in addition to the standard profession class. This allows us to study how specific demographics affect the classification accuracy.

Finally, we also explore and analyze the per-class accuracy results when classifying the specified demographic, as discussed in later sections.

4.3 Top-1 Results

The top-1 results when classifying two out of the 5 demographics are shown in Table 3 (Additional results for the other demographics will be included in Section 10).

CLIP performs fairly well when classifying profession, which was to be expected because CLIP’s training data likely includes richer cross-modal representations related to profession. As a result, all of the accuracies are quite high. Despite this already high performance, D3G is able to still improve performance. This implies that providing diverse demographics can still improve CLIP’s understanding of well-established concepts.

We are able to gain a better understanding of D3G’s efficacy when we look at the performance gains on Race 7. For this demographic, CLIP performance is much worse. Once again, the model performs better when the prompt contains information regarding profession, due to the increased likelihood of this information within the training

data; however, when this information is omitted, the performance on the other demographics is abysmal. With the accuracy only scoring around 10-15% more than random guessing (which should be around 14% top-1 accuracy), this shows that CLIP does not have a deep understanding of race and other demographics.

With this in mind, by simply implementing D3G, we are able to push the accuracies up to by 4-7%. This indicates that coupling the standard embeddings with diverse data that has been generated, improves CLIP’s understanding of concepts that were previously misunderstood. In addition, it is important to note that Average Image D3G typically performs better than the standard method. Once again, this makes sense and conforms to our hypothesis. Generating diverse data pushes the embeddings closer to the ground-truth position within embedding space, resulting in more accurate predictions for classes the model may not fully understand.

These results are highly promising, and we can learn a bit more about the effect of D3G on these results by looking at the weights utilized to produce these scores.

4.4 Weighting Strategy

Recall that in order to classify a query image, we form a weighted sum of the embeddings between the text prompt, and the generated images. The ratio of text-to-image weighting is dictated by scanning values until an optimal state is found. This is necessary, because for certain images, the text embeddings will contribute more to the classification result than the image embeddings and vice versa. With this in mind, the weighting ratio between images and text is also an indicator of how much the generated images from D3G actually help given a defined demographic. Knowing

Demographic	Prompt	White	Black	Latino	East Asian	South East Asian	Indian	Middle Eastern
Race 7	Profession	68.19	70.90	15.38	43.46	20.59	57.58	13.80
	Race 7	8.92	66.42	11.54	52.74	8.82	60.61	32.68
	Race 4	-	-	-	-	-	-	-
	Gender	18.07	73.13	11.54	35.02	17.65	51.52	34.93
	Age	12.29	68.66	11.54	43.46	14.71	69.70	29.58

Table 5: **Standard D3G** per-class results when classifying the specified demographic. Note that all of the prompts are as described in Table 2 (e.g. "A photo of a black person", or "A photo of a 30-39 year old doctor"). More examples are provided in Section 10

this information, we can then start to understand exactly what our results mean in the broader scope.

When viewing the weights from Table 4, we see the same trends that were displayed within Table 3, but we get a glimpse as to why D3G had minimal performance gains. When classifying profession, most of the text weights for Standard D3G are quite high, being roughly around 85-90%; however, whenever see larger increases in accuracy for D3G, we also see an increased weighting of the generated images. This is especially evident when classifying race 7. Once again, the prompts that utilized professions were able to get somewhat higher accuracies, due to the structure of the dataset; however, for every other race 7 evaluation, the generated images played a major role in the classification results. The fact that images were consistently weighted around 30% shows that the diversity matters when classifying demographics.

The results analyzed from our top-1 results and their corresponding weighting strategies, show *that* the method works; however, the per-class results give us a deeper understanding of *why* the method works.

4.5 Per-Class Results

For these results, we primarily reference Table 5; however, note that additional per-class results are included within Section 10. When classifying race 7, we know that the best performance gains were from including gender and age into the prompts. Focusing on these rows, we can see a few interesting trends. For instance, including information about gender improves the accuracy for Black and Middle Eastern people the most. This is likely due to the fact that within CLIP's training data, these populations have gender underrepresented. Within Section 5.2, we will later discuss future methods of confirming this hypothesis.

Now that we understand which demographics help classification accuracies, we can now start to extend these inferences across demographics. Images and text related to East Asian people likely did not have rich cross-modal representations because race 7 helped the most for this demographic. This means that simply generating images of diverse races was able to significantly boost the accuracy. Similarly, age was the most useful demographic when classifying people in images as Indian. This was quite

surprising, and as we discuss in Section 5.2, we intend to further explore the impact of these results by including additional metrics such as precision, recall, specificity, and F1 score.

Another trait of these per-class results emerges when we compare the accuracy ratios across demographic columns. For instance, Black generally achieves a higher per-class accuracy than all of the other demographics, with Indian and East Asian obtaining the second and third highest overall per-class accuracies across all of the prompts. Alongside this, Latino, South East Asian, and White achieve some of the lowest per-class accuracies overall across all prompts. We were very surprised by this outcome, especially by the fact that race 7 was the worst performing prompt for White, which had the majority representation within the dataset. Intuitively, this my imply that providing diverse representations can also move embeddings away from the correct position in embedding space. In order to combat this, we may be able to strategically weight generated image and text prompt embeddings in relation to their demographic proportions within the dataset (e.g. if Latino is underrepresented within the dataset, then we will up-weight the Latino embeddings). This idea is further explored within Section 5.2.

Finally, we did not describe the results for profession, due to the fact that we cannot infer why these demograpics performed best, due to the fact that CLIP leverages profession information to make it's predictions, but the dataset is catered towards profession. This means that the increased accuracies could be either due to the profession information within the prompt, or the images generated of each profession. Either way, we will need to run more tests to fully understand this. We intend to evaluate on other datasets, so we can understand whether this correlations indication causation; however, these are very promising results.

5 Discussion

5.1 Assumptions

Within this paper, wwo prominent assumptions made:

1. The generative model has a better learned representation of the true distribution of the data (due to its increased complexity and data diversity).
2. The base multimodal model can distinguish between

similar classes. Our method will not improve performance if this is not the case.

These assumptions are necessary for D3G to function properly, but they are not unreasonable for a zero-shot setting. The generative model must have a better learned representation of the true data distribution, because it needs to be able to generate images that accurately represent the desired concept. If the model cannot generate useful images, then D3G will revert to using the baseline CLIP method, with text-based classification.

In addition, we need our base model to be able to distinguish between similar classes, because if two classes correspond to the same point within embedding space, then they our model cannot distinguish them. Similarly, we need this assumption so that the weighted sum of the image and text embeddings actually pushes the embedding towards the true embedding, and not just in a random direction. If the base model couldn't distinguish between certain classes, then we would have no guarantee that creating a weighted sum actually improves classification, because the model would be completely guessing in that case. In the future, we may be able to validate this assumption by comparing the embeddings within embedding space to ensure they are an adequate distance apart, but for now this will be maintained as an assumption.

These assumptions on their own are not unreasonable; however, in certain circumstances they may become limitations as discussed later.

As mentioned previously, this research is crucial because models such as CLIP, dictate are frequently used to filter large datasets such as DataComp-1B or LAION-2B. If CLIP performs so poorly when classifying demographics, then these biases will be reinforced on all models trained with the datasets. This issue has compounding effects, and so to reduce demographic bias within image generation, object detection, and content moderation models, we **must** start with image classification.

5.2 Future Work

With such promising results from this project, there are many steps we intend to take in the future, in order to ensure this method is as robust as possible.

To start, we aim to include additional metrics that properly quantify the balance between demographics to better understand how D3G balanced the predictions of the multimodal classifier. We specifically hope to investigate the robustness of our approach to class imbalance, data redundancy, and noise levels.

For this paper, we decided to simply average the embeddings of all images generated with D3G, however this may not be the most effective process. Even though we generate images of a diverse range of demographics, these demographics are not weighted equally by CLIP (as

demonstrated previously in Table 5), due the training data. This means that by utilizing the CLIP image encoder to get embeddings for all of our images, we are only offsetting the existing bias but this does not create a neutral embedding; rather, it creates an embedding that still emphasizes the existing bias but is slightly more balanced across demographics. In order to combat this, we aim to explore how we can create a weighted sum of the embeddings from individual images, that is informed by the demographics of the training data and of the broader world. Intuitively, if CLIP tends to favor one demographic, then we will down weight those images, and vice-versa if CLIP rarely selects another demographic. In this way, we can robustly enforce equity within CLIP's predictions.

In addition to this step, in the future we also aim to utilize OpenCLIP so that we can accurately draw conclusions about the model's predictions in relation to the training data. Since we solely used CLIP as a baseline for this paper, we are unable to confidently state that the distribution of the training data led to the model's sometimes biased predictions; however, this is strongly implied. By utilizing a model with open training data and architecture, we can draw these conclusions with certainty. Researchers are starting to explore demographic bias within LAION-2B and DataComp-1B (the training data for certain OpenCLIP models), and we aim to leverage this knowledge for future implementations.

We would also like to expand our evaluation suite to multiple datasets. Currently we only evaluate on 2,000 images from IdenProf, but we could start by utilizing the full dataset of 11,000 training and test images, since we are not training and we want a wider pool of images. In addition, we intend to perform similar tests over the FairFace dataset. These results would more effectively isolate CLIP's capabilities predicting the demographics outlined in this paper, since the FairFace dataset was constructed with these demographics in mind. This is especially important, because we found that CLIP was able to leverage the semantic information regarding professions within the dataset, in order to classify race 7 more accurately. By removing professions as a factor, we will be able to fully explore CLIP's performance on such tasks.

An important note, was that we were particularly intrigued by CLIP's inadequate performance when classifying demographics such as race 7, so we also aim to conduct an analysis on the individual classification results, combined with metrics such as precision, recall, specificity, and overall F1 score in order to better understand whether CLIP's performance on these demographics is statistically significant. If the positive predictions are solely informed by demographic stereotypes, then we aim to expose these weaknesses and combat them with D3G.

Finally, in addition to generating images based on the demographics, we also aim to explore methods of retrieving images, or modifying the demographics of the query

image in-place. Modifying the existing query image to get diverse demographics, may reduce the impact of stereotypes enforced by the image generation model, and result in classifications that are much more accurate.

6 Conclusion

Image classification remains a challenging task despite advancements in multimodal models like CLIP that leverage semantic similarities across vision and language. Low-capacity models often suffer from underfitting, leading to poor performance; however, the generation of high-quality data with rich cross-modal representations is also difficult. Imbalanced demographics in datasets can cause predictions to bias toward more represented classes, pushing those who are underrepresented to the wayside. Our study highlights these issues and their impact on zero-shot image classification, proposing Diverse Demographic Data Generation (D3G) as a solution. This training-free, zero-shot method enhances classification accuracy and reduces demographic bias in pre-trained multimodal models by providing diverse demographic data at inference time, demonstrating improved performance for these models.

7 Ethics Statement

The fact that we are utilizing image generation models for D3G provides significant potential for negative societal impact. For instance, the images generated by the model can often reinforce certain demographic biases. This is to be expected, because the prompts used within this paper are quite vague; however, this also shows that the generative model has learned visual stereotypes from its training data. The stereotypes within the generated images is why they should only be used as a weighted sum with the text, and never as the sole ground-truth signal. Up-weighting the images too much, provides opportunity for unethical image generations.

One potential way to combat this issue of stereotypes within generation, is to utilize the method discussed in Section 5.2, where we modify the query image in-place in order to reduce the room for error, while still increasing demographic diversity.

Along with this, our use of generative modelling allows for potentially unethical prompting. The only restrictions on prompting are those enforced by Stable Diffusion XL; however, due to the open-source nature of the model, many of these restrictions can be circumvented. We do not condone the use of D3G to generate and hateful, demeaning, or otherwise unethical data. This method should only be used within appropriate contexts, and primarily as a means of increasing pre-trained model diversity ad hoc.

The selection of demographics used within our classification process was mainly a result of the process used to create the FairFace dataset (Karkkainen and Joo 2021). The authors defined the races used to be based on commonly accepted race classification from the U.S. Census Bureau; however, we acknowledge that does not properly represent

the racial landscape of the world. It is important to note that the authors decided to use skin color as a proxy to race, combined with annotations about physical attributes. This means that the annotations used to construct the dataset and train the FairFace classification model used to create labels for IdenProf, may contain annotator bias. This is evident in the gender demographic. The authors mentioned it would be impossible to perfectly balance the gender predictions of their model, outside of a lab setting. Finally, ages were simply segmented into common age groups. The decision to use these demographic categories limits the conclusions we can draw in this paper, regarding the impact of all relevant demographics to classification accuracy.

Finally, D3G is a technique that *does not* remove demographic biases, but rather, it offsets learned biases. This means that the method can either reduce or accentuate human bias, and should not be used as a universal architecture to improve multimodal model fairness and accuracy. If the images generated contain harmful bias, then this technique could make the performance worse and much more inequitable.

8 Limitations

Due to this paper being focused on classification, a significant limitation is with regards to demographic intersectionality. People that fit into multiple demographics within the same category (i.e. people who are biracial), will suffer from only being classified as a single demographic. This is a limitation, because it is a known issue that cannot be surmounted using standard metrics within image classification. Future methods may be able to explore intersectionality by retrieving the top-k classified demographics; however, this would be difficult in a zero-shot setting, where no additional information about the query image is provided.

A second major limitation is the fact the D3G can only perform well if the pretrained models are able to effectively distinguish between the demographics being classified. As mentioned within previously, if the multimodal model embeds two demographics to the same point in embedding space, or if the image generation model cannot generate good images for a given demographic, the technique will fail. This is typically not an issue for the broad demographics covered within this paper; however, it may become more difficult as the classes become more fine-grained.

A final limitation is the fact that D3G utilizes pre-trained models for every step of the pipeline. This partially is also the most useful part of the technique; however, it also means that the limitations of the pretrained models will extend to D3G. The abilities or inabilities of the generative model will result in the final classification accuracies. Similarly, the quality of the embeddings produced from the multimodal model will dictate the effect D3G will have on classification accuracy.

9 Individual Contributions

I solely completed this full project. Within this paper I utilize the “we” subject pronoun to maintain standard academic paper writing practices.

10 Appendix

References

- Breiman, L. 1996. Bagging predictors. *Machine learning*, 24: 123–140.
- Gadre, S. Y.; Ilharco, G.; Fang, A.; Hayase, J.; Smyrnis, G.; Nguyen, T.; Marten, R.; Wortsman, M.; Ghosh, D.; Zhang, J.; et al. 2024. Datacomp: In search of the next generation of multimodal datasets. *Advances in Neural Information Processing Systems*, 36.
- Karkkainen, K.; and Joo, J. 2021. 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*, 1548–1558.
- Podell, D.; English, Z.; Lacey, K.; Blattmann, A.; Dockhorn, T.; Müller, J.; Penna, J.; and Rombach, R. 2023. Sdxl: Improving latent diffusion models for high-resolution image synthesis. *arXiv preprint arXiv:2307.01952*.
- Pratt, S.; Covert, I.; Liu, R.; and Farhadi, A. 2023. What does a platypus look like? generating customized prompts for zero-shot image classification. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, 15691–15701.
- Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; et al. 2021. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763. PMLR.
- Ramesh, A.; Pavlov, M.; Goh, G.; Gray, S.; Voss, C.; Radford, A.; Chen, M.; and Sutskever, I. 2021. Zero-shot text-to-image generation. In *International conference on machine learning*, 8821–8831. Pmlr.
- Rombach, R.; Blattmann, A.; Lorenz, D.; Esser, P.; and Ommer, B. 2022. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 10684–10695.
- Udandarao, V.; Gupta, A.; and Albanie, S. 2022. Sus-x: Training-free name-only transfer of vision-language models. *arXiv preprint arXiv:2211.16198*.
- Wallingford, M.; Ramanujan, V.; Fang, A.; Kusupati, A.; Mottaghi, R.; Kembhavi, A.; Schmidt, L.; and Farhadi, A. 2024. Neural priming for sample-efficient adaptation. *Advances in Neural Information Processing Systems*, 36.
- Yang, Z.; Luo, T.; Wang, D.; Hu, Z.; Gao, J.; and Wang, L. 2018. Learning to navigate for fine-grained classification. In *Proceedings of the European conference on computer vision (ECCV)*, 420–435.

Demographic	Prompt	Text
Race 7	"A photo of a <race> person"	A photo of a black person
Race 4	"A photo of a <race> person"	A photo of a black person
Profession	"A photo of a <race> <prof>"	A photo of a black doctor
Gender	"A photo of a <race> <gender>"	A photo of a black male
Age	"A photo of a <age> year old <race> person"	A photo of a 30-39 year old black person

Table 6: Example diverse demographic texts for classifying *race 7*. The texts for classifying race 4 will be identical, just fewer racial classes.

Demographic	Prompt	Text
Gender	"A photo of a <gender>"	A photo of a female
Profession	"A photo of a <gender> doctor"	A photo of a female doctor
Race 7	"A photo of a <race> <gender>"	A photo of a black female
Race 4	"A photo of a <race> <gender>"	A photo of a black female
Age	"A photo of a <age> year old <gender> person"	A photo of a 30-39 year old female

Table 7: Example diverse demographic texts for classifying *gender*.

Demographic	Prompt	Text
Age	"A photo of a <age> year old"	A photo of a 30-39 year old
Profession	"A photo of a <age> year old doctor"	A photo of a 30-39 year old doctor
Race 7	"A photo of a <age> year old <race> person"	A photo of a 30-39 year old black person
Race 4	"A photo of a <age> <race>"	A photo of a 30-39 year old black person
Gender	"A photo of a <age> <gender>"	A photo of a 30-39 year old female

Table 8: Example diverse demographic texts for classifying *gender*.

Demographic	Prompt	White	Black	Latino	East Asian	South East Asian	Indian	Middle Eastern
Race 7	Profession	68.92	70.90	15.38	43.46	20.59	57.58	13.24
	Race 7	9.40	61.94	15.38	66.24	5.88	60.61	26.48
	Race 4	-	-	-	-	-	-	-
	Gender	20.72	67.91	15.38	44.73	11.76	51.52	29.86
	Age	11.57	65.67	11.54	59.92	14.71	69.70	25.07

Table 9: *Average D3G* per-class results when classifying the specified demographic. Note that all of the prompts are as described in Table 2.

	Prompt	Chef	Doctor	Engineer	Farmer	Firefighter	Judge	Mechanic	Pilot	Police	Waiter
Profession	Profession	95.54	98.80	97.62	98.36	94.12	98.20	86.49	99.37	100.0	84.34
Profession	Race 7	94.27	98.80	96.43	98.36	96.08	99.40	81.08	98.74	100.0	84.94
Profession	Race 4	94.90	98.80	96.43	98.36	96.08	99.40	82.43	98.74	100.0	84.34
Profession	Gender	96.82	98.80	96.43	96.72	98.04	99.40	90.54	99.37	100.0	87.35
Profession	Age	97.45	98.19	96.43	98.36	98.04	98.80	85.14	97.48	100.0	80.72

Table 10: *Standard D3G* per-class results when classifying the specified demographic. Note that all of the prompts are as described in Table 2. These results for race 7 are shown in Table 5.

	Prompt	Chef	Doctor	Engineer	Farmer	Firefighter	Judge	Mechanic	Pilot	Police	Waiter
Profession	Profession	93.63	98.80	98.81	100.0	94.12	96.41	83.78	99.37	100.0	90.36
Profession	Race 7	93.63	98.80	97.62	98.36	94.12	97.60	81.08	98.74	100.0	90.36
Profession	Race 4	93.63	98.80	96.43	98.36	94.11	98.80	82.43	98.74	100.0	87.35
Profession	Gender	96.18	98.80	96.43	98.36	98.04	99.40	90.54	99.37	100.0	89.16
Profession	Age	94.90	98.80	95.24	100.0	94.12	96.41	82.43	95.60	100.0	92.77

Table 11: *Average D3G* per-class results when classifying the specified demographic. Note that all of the prompts are as described in Table 2