WILL THE INCLUSION OF GENERATED DATA AMPLIFY BIAS ACROSS GENERATIONS IN FUTURE IMAGE CLASSIFICATION MODELS?

***Zeliang Zhang, *Xin Liang,* Mingqian Feng, Susan Liang, Chenliang Xu**
University of Rochester, NY

ABSTRACT

As the demand for high-quality training data escalates, researchers have increasingly turned to generative models to create synthetic data, addressing data scarcity and enabling continuous model improvement. However, reliance on self-generated data introduces a critical question: *Will this practice amplify bias in future models?* While most research has focused on overall performance, the impact on model bias, particularly subgroup bias, remains underexplored. In this work, we investigate the effects of the generated data on image classification tasks, with a specific focus on bias. We develop a practical simulation environment that integrates a self-consuming loop, where the generative model and classification model are trained synergistically. Hundreds of experiments are conducted on Colorized MNIST, CIFAR-20/100, and Hard ImageNet datasets to reveal changes in fairness metrics across generations. In addition, we provide a conjecture to explain the bias dynamics when training models on continuously augmented datasets across generations. Our findings contribute to the ongoing debate on the implications of synthetic data for fairness in real-world applications.

1 INTRODUCTION

As models continue to evolve and become more sophisticated, the demand for large amounts of high-quality training data has escalated (Alzubaidi et al., 2023). Traditionally, web data has been the primary resource for enhancing model performance (Deng et al., 2024). However, as this source becomes fully exploited, researchers have begun to explore alternative methods. One promising approach is to leverage generative models to create synthetic data (Fan et al., 2024; Meng et al., 2022; Zhou et al., 2023; Yang et al., 2023), thereby fueling continuous training cycles, as shown in fig. 1. This innovative self-sustaining pipeline effectively mitigates the issue of data scarcity, allowing models to improve iteratively with the help of their own generated outputs (Chen et al., 2024; Lu et al., 2024). Despite the apparent advantages, this strategy introduces a crucial and complex debate: *Will the reliance on self-generated data eventually lead to model degradation*?

Some research has attempted to answer this question. On the one hand, Azizi et al. (2023) and Zhou et al. (2023) use the diffusion model to generate synthetic image data to augment the training set and observe the performance improvement in image classification tasks. Zheng et al. (2024) analyze the positive impact of generative data enhancement on small-scale datasets from a theoretical perspective. Hammoud et al. states that a carefully designed generated data augmentation strategy could be helpful to alleviate the long tail problem. On the other hand, Alemohammad et al. (2024) show that purely adding the generated data to agent training could eventually cause model

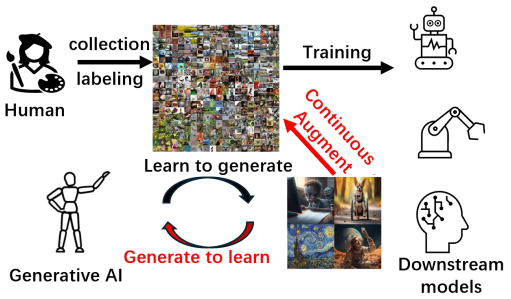


Figure 1: Generative models can be leveraged to generate more data to augment the training set, then help the downstream models training.

degradation, with their quality or diversity progressively decreasing. Singh et al. (2024) demonstrate that the use of synthetic data could cause a large performance drop in model robustness. The debate is still ongoing and remains unsettled.

It is important to note that although many research efforts are put into analyzing the influence of generative data on overall model performance, few of them explore the impact of generative data on the model bias, especially on the model's behavior in the worst-performing subgroups. Previous work (Zhang et al., 2024) has identified that models often behave significantly differently across various unknown subgroups, showing the critical role of model fairness in real-world applications. In the context of generative data, we raise a new question in this paper: will the inclusion of generated data help alleviate the model bias problem, or could it potentially make it worse? This question and its answer are significantly connected with other bias issues, *e.g.*, demographic parity (Loukas & Chung, 2023), equalized odds (Grant, 2023), maximum disparity (Roh et al., 2020), spurious correlation (Seo et al., 2022).

Intuitively, the bias issue is probably to be amplified because generated data is increasingly leveraged in training models across successive generations. Previous findings (Sehwag et al., 2022; He et al., 2024) reveal that generative models tend to sample data from high-density regions, leaving low-density data heavily under-explored. This imbalanced sampling introduces a natural skew in the dataset used to augment training, thereby exacerbating the bias present in the model. However, is this assumption accurate? To the best of our knowledge, no research has thoroughly explored this question. This lack of exploration leaves a critical issue unresolved, potentially creating an unknown risk in practical applications.

In this work, we study the impact of generated data on the model bias through the lens of the image classification problem, one of the most fundamental tasks of computer vision and deep learning. Our approach differs from previous studies in two key aspects: First, we focus on the impact of generated data on the model bias. Second, we create a more practical simulation environment by building a self-consuming loop that trains the generative model and the image classification model synergistically. We conduct experiments on three datasets, including colorized MNIST (Kim et al., 2019a), CIFAR-20/100 (Zhang et al., 2024), and Hard ImageNet (Moayeri et al., 2022b), to observe and analyze changes in various fairness metrics.

We summarize our contributions and key findings as follows:

1. We design and implement a scalable, self-consuming simulation environment. Our method interleaves dataset augmentation and model training across different generations.
2. We introduce data stacking and expert-guided filtering approaches to overcome data explosion and inconsistent data quality issues.
3. We conduct extensive experiments on three popular datasets to examine and reveal the impact of cross-generation generated data on model performance and bias.
4. We systematically analyze the factors causing diverse model bias behaviors.

2 Related work

2.1 Generative model and its application

Generative models have become a cornerstone of modern machine learning, particularly in the domain of data augmentation and synthetic data generation (Akkem et al., 2024). Early approaches, such as Generative Adversarial Networks (GANs) (Goodfellow et al., 2020), revolutionized the field by enabling the creation of highly realistic synthetic data through a process of adversarial training between a generator and a discriminator. More recently, diffusion models (Croitoru et al., 2023) have gained prominence due to their ability to generate high-quality data through a denoising process, offering an alternative to traditional GAN-based approaches. These generative models have been widely adopted in various tasks, including image synthesis (Liao et al., 2020), text generation (Li et al., 2018), and data augmentation (Antoniou et al., 2017), proving their efficacy in improving model performance. In this work, we leverage two generative models, including the conditional GAN and text-to-image diffusion.

The advent of generative models has significantly expanded the possibilities for data augmentation by enabling the creation of entirely new data samples that mimic the distribution of the original dataset. For instance, Azizi et al. (2023) and Zhou et al. (2023) leverage diffusion models to generate synthetic images, successfully augmenting training sets and improving classification accuracy. Similarly, Zheng et al. (2024) explore the theoretical underpinnings of generative data augmentation, particularly in the context of small-scale datasets. However, the impact of using synthetic data is not without its challenges. Alemohammad et al. (2024) highlight that indiscriminate inclusion of generated data in training can lead to model degradation, where the model's performance deteriorates as the quality and diversity of the generated data decrease over time. Hammoud et al. observe a related phenomenon, noting that a carefully designed strategy for data augmentation could mitigate issues such as the long-tail problem. Further, Singh et al. (2024) demonstrate that the use of synthetic data can significantly undermine model robustness, leading to performance drops.

2.2 Bias in deep learning models

Many efforts have been made on the model bias. Kotek et al. (2023) investigate the behavior of large language models on gender bias. Liu et al. (2022) measure the political bias in language models. Zhang et al. (2024) identify the existence of subgroup bias in image classifiers. Hosseini et al. (2018) find the shape bias learning by convolutional neural networks. Khayatkhoei & Elgammal (2022) discover generative models can easily learn the spatial bias from the data. Heinert et al. (2024) and Hönig et al. (2024) research on texture bias in deep learning models.

There are also many fairness metrics to help evaluate bias (Kim et al., 2019b; Lin et al., 2022). Important fairness metrics include demographic parity (Jiang et al., 2022), which ensures that positive classification rates are equal across different demographic groups, and equalized odds (Romano et al., 2020), which requires that true positive and false positive rates are consistent across groups. Equal opportunity (Wang et al., 2023) further emphasizes equal true positive rates, ensuring that no group is disadvantaged in correct classifications.

3 Generate to Learn: Building a Scalable and Self-Sustaining Simulation Environment

3.1 A Simple yet Practical Framework for Simulation

To better understand the impact of the generated data on future model training, we design and implement a simulation environment grounded in real-world practices. The environment comprises four core components: subgroup construction, base model initialization, dataset augmentation, and future model development.

∘ *Subgroup Construction*. Our environment is designed to study the effects of generated data on model bias, making it essential to establish clear and practical attributes for bias evaluation. Inspired by Zhang et al. (2024), we manually partition the original dataset into multiple subgroups, where subgroups within the same class share similar semantics. The introduction of bias is controlled by adjusting these subgroup partitions. During the training process, the models remain unaware of the subgroup partitions, which are only revealed during the evaluation stage to assess model bias.

∘ *Base Model Initialization*. We construct and randomly initialize a base generative model $g(\cdot)$. This model is then trained from scratch on the dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{N}$, where x represents the sample to generate, y is the corresponding label, and N is the number of training samples in $\mathcal{D}$. The model is trained until it converges sufficiently on $\mathcal{D}$.

∘ *Dataset Augmentation*. Once the base model is initialized, we use the generative model $g(\cdot)$ to generate data that approximates the distribution of the training dataset $\mathcal{D}$, thereby augmenting the original dataset. Because previous study (Zheng et al., 2024) has shown that training exclusively on generated data can eventually cause model failures, we adopt an alternative strategy (Azizi et al., 2023; Zhou et al., 2023), mixing the original data with generated data at a ratio of $p\%$.

∘ *Future Model Development*. In addition to the base generative models, our task involves two types of models: downstream models and subsequent generative models. The downstream model corresponds to an image classification model optimized with cross-entropy loss. The generative model

is a re-initialized version of $g(\cdot)$, trained on the augmented dataset from the previous generation. Unlike previous similar work that considers only a single generation, we incorporate generated data from multiple continuous generations, creating a more realistic and practical scenario.

We leverage the above core components to build our simulation environment. We begin with *subgroup construction* to study the model behaviors of interest. At each generation, we *(re)initialize the base model* using the current dataset, which may have been augmented. This model is then used to generate additional data for *dataset augmentation*. Finally, the *downstream models are developed* on the dataset augmented by the current-generation generative model.

3.2 Scaling the Simulation for Real-World Practice

Two significant challenges remain in our environment, limiting its scalability for simulating real-world practices: 1) *Data Explosion*: As the number of generations increases, the volume of generated data grows continuously, leading to a significantly larger training set and resulting in unbearable training time consumption. 2) *Inconsistent Data Quality*: Due to the inherent uncertainty in the generative process, the quality of data produced by the generative model across different generations may vary, potentially leading to degradation in the performance of future models.

We propose two strategies to incorporate into our simulation environment to address these challenges, including the *Data Stacking* and *Expert-guided Filtering*.

○ *Data Stacking*. We maintain a first-in-first-out queue to store the generated data. Specifically, we set the capacity of the queue to D. We continuously use the updated generative model to generate data with a volume of S and fill the queue until it reaches capacity, *i.e.*, the maximum number of generations that can be accommodated is D/S. Once the queue is full, the oldest data will be removed to make space for newly generated data.

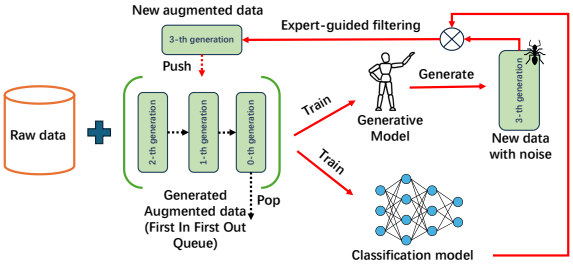


Figure 2: We continuously leverage new generators to produce additional images that enhance the training process, employing data stacking and expert-guided filtering to maintain high quality. We highlight the trajectory of the self-consuming loop in red.

○ *Expert-Guided Filtering*. We introduce two expert-guided strategies to filter low-quality samples and improve the quality of the training set. The first strategy involves conducting a human study to score the generated samples and removing those that are easily recognized as generated content. The second strategy leverages the CLIP and our trained classification model of the last generation to score the generated samples based on the prediction uncertainty (Gal & Ghahramani, 2016), filtering out the bottom $r\%$ based on their scores.

3.3 Evaluation Metrics for Assessing Model Bias

It is important to evaluate model performance, including bias, during the development process across generations. Consider an input $\boldsymbol{x} \in \mathcal{X}$ from the initial meta training dataset in our simulation environment, associated with a ground-truth label $y \in \mathcal{Y}$. Assume the dataset comprises L distinct classes, so $\mathcal{Y} = 1, 2, \ldots, L$. We hypothesize that each class is further divided into G subgroups, assuming for simplicity that each class contains an equal number of subgroups, resulting in a total of $L \times G$ subgroups across the dataset. For each input $\boldsymbol{x}$, its subgroup membership is denoted by $g \in 1, 2, \ldots, G$. The existence of such unknown subgroups and the varying model performance across these subgroups contribute to the presence of model bias. In our environment, we use several criteria to evaluate model performance, including overall performance, multi-group equality of opportunity, multi-group disparate impact, maximum disparity, and sub-population performance. Among these metrics, we extend the conventional equality of opportunity and disparate impact to the content of image classification task with multiple tasks and multiple unknown attributes.

Overall Performance. We use the Fréchet inception distance (FID) (Heusel et al., 2017) and the classification accuracy (Acc) as metrics for the evaluation of the generative model and the downstream

classification model.

$$\text{FID}_n = \|\mu_c - \mu_g\|_2^2 + \text{Tr}(\Sigma_c + \Sigma_g - 2(\Sigma_c\Sigma_g)^{\frac{1}{2}}), \;\; \text{Acc}_n = P\left(f_n\left(x\right) = y_x\right), \tag{1}$$

where μ_c and Σ_c are the mean and variance matrix of the feature vector extracted from Inception-V3 (Szegedy et al., 2015) on the original clean samples, μ_g and Σ_g are those from the generated samples, y_x is the ground-truth associated with the sample x, and n indicates the number of generation.

Equality of Opportunity. The equality of opportunity (Ferreira & Peragine, 2013) measures whether every subgroup is treated equally by the model under study. In our simulation environment, we compute the equality of opportunity (EO) under the background of multiple groups as follows:

$$\text{EO}_n = 1 - \frac{1}{\binom{G}{2}} \sum_{i,j<G, i\neq j} \left\|\text{TPR}_n^i - \text{TPR}_n^j\right\|, \tag{2}$$

where we denote TPR_n^i as the the true positive rate of i-th subgroup in the n-th generation. It can be computed as $\text{TPR}_n^i = P\left(f_n(\boldsymbol{x}) = y \mid y = y_x, g = i\right)$, indicating the probability that the model f_n correctly classifies an input $\boldsymbol{x}$ from the i-th subgroup with the ground-truth label $y = y_x$.

Disparate Impact. Disparate impact (Feldman et al., 2015) measures whether different subgroups receive positive outcomes at similar rates. In our simulation environment, we extend this concept to multiple groups, defining the multi-group disparate impact (DI) as follows:

$$\text{DI}_n = 1 - \frac{1}{\binom{G}{2}} \sum_{i,j<G, i\neq j} \left\|\frac{P(f_n(\boldsymbol{x}) = y_x \mid g = i)}{P(f_n(\boldsymbol{x}) = y_x \mid g = j)} - 1\right\|, \tag{3}$$

where $P(f_n(\boldsymbol{x}) = y_x \mid g = i)$ denotes the probability that the model f_n assigns a positive outcome (e.g., $y = y_x$) to an input $\boldsymbol{x}$ from the i-th subgroup.

Maximum Disparity. Maximum disparity measures the largest difference in model performance between any two subgroups. We compute the maximum disparity (MD) as follows:

$$\text{MD}_n = \max_{i,j<G, i\neq j} \left\|\text{TPR}_n^i - \text{TPR}_n^j\right\|. \tag{4}$$

Subgroup Performance. In addition to the aforementioned metrics for single-bias evaluation by pair-wise computation, we evaluate model performance by examining the accuracy of the multiple worst-performing subgroups. For each superclass c, we calculate the accuracy of its G subgroups, denoted as $\text{Acc}_{c,g}$, and sort these accuracies in ascending order, $\text{Acc}_{c,(1)} \leq \text{Acc}_{c,(2)} \leq \cdots \leq \text{Acc}_{c,(G)}$. We then compute the average accuracy for each rank k across all superclasses:

$$\overline{\text{Acc}}_{(k)} = \frac{1}{C} \sum_{c=1}^{C} \text{Acc}_{c,(k)}, \tag{5}$$

where C is the total number of superclasses. This allows us to assess the model's performance across the most challenging subgroups.

Among these metrics, MEO (eq. (2)), DI (eq. (3)), and MD (eq. (4)) assess single-bias evaluation, and subgroup performance evaluates (eq. (5)) the impact of multiple biases.

***Why do we select these metrics*?** We do not choose to use the one-vs-rest (OvR) strategy (Jung et al., 2021) for evaluating fairness metrics in our multi-class classification tasks because OvR reduces multi-class problems to multiple binary subproblems, potentially missing the intricate biases and class interactions inherent in genuine multi-class contexts, thus overlooking unfairness arising from these interactions (Friedler et al., 2019). Additionally, OvR could introduce significant data imbalance in each binary subproblem, especially when class distributions vary greatly, which adversely affects classifier performance and distorts fairness metrics, leading to unreliable evaluations (Brzezinski et al., 2024). Instead, we employ fairness metrics specifically designed for multi-class classification — MEO, DI, and MD — to assess fairness across all classes simultaneously, preserving the integrity of the multi-class problem and providing a more accurate evaluation (Mazijn et al., 2021). This approach ensures that our fairness assessments reflect the complexities of multi-class classification, effectively manage potential data imbalances, and align with our objective to enhance fairness in a comprehensive and contextually appropriate manner.

Table 1: Evaluation of FID across generations for different generative models trained on various datasets, including colorized MNIST w/wo bias initialization, CIFAR-20/100, and Hard ImageNet. The 1st generative model is trained on the original dataset without the inclusion of generative data.

Dataset	Initialization	Number of generations									
		1	2	3	4	5	6	7	8	9	10
Colorized MNIST	Unbiased	111.7	108.9	107.8	107.1	104.5	103.6	101.0	100.5	103.6	106.3
	Biased	109.4	106.0	107.03	106.4	105.3	114.2	109.0	108.6	109.1	116.5
CIFAR-20/100	N/A	249.3	213.6	210.6	217.7	218.8	224.9	233.1	226.5	220.6	223.0
Hard ImageNet	N/A	57.6	50.9	186.4	118.8	-	-	-	-	-	-

4 EXPERIMENTS

4.1 EVALUATION SETUP

Datasets. We studied three datasets: Colorized MNIST, CIFAR-20/100, and Hard ImageNet. ① The Colorized MNIST dataset is a modified version of the original MNIST (LeCun, 1998), where three colors—red, blue, and green—are added to the images. We created two versions of this dataset. In the first, the three colors are uniformly applied across different classes. In the second, the colors are applied with uneven ratios, introducing a bias in the color distribution. ② The CIFAR-20/100 dataset is derived from CIFAR-100 (Alex, 2009) by grouping every five subclasses with similar semantic meaning into one single superclass, resulting in 20 classes. ③ Hard ImageNet (Moayeri et al., 2022b), a challenging subset of the ImageNet dataset(Deng et al., 2009), consists of 15 classes and contains various spurious correlations that can undermine the reliability of models trained on it.

Models. In the experiments with colorized MNIST and CIFAR-20/100, we consider five models: LeNet (LeCun et al., 1998), AlexNet (Krizhevsky et al., 2012), VGG-19 (Simonyan & Zisserman, 2014), ResNet-50 (He et al., 2016), and MobileNet-V3 (Howard et al., 2019). For the Hard ImageNet experiment, we exclude the smallest model, LeNet, and additionally include a larger model, DeiT-S (Touvron et al., 2021). These models are sourced from the PyTorch library (Paszke et al., 2019), with the final layer modified to fit the specific classification tasks. We use GANs (Radford, 2015) to learn and generate the colorized MNIST and CIFAR-20/100 datasets, while stable-diffusion-1.5 (Rombach et al., 2022) is employed for generating the Hard ImageNet dataset.

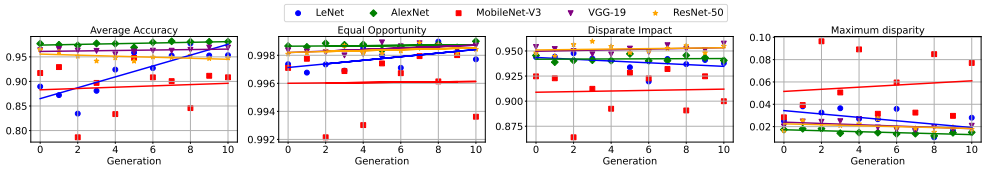
Metrics. We evaluate model performance across all datasets based on classification accuracy. For the Colorized MNIST and CIFAR-20/100 datasets, which have explicit subgroups but are trained only at the superclass level, we also assess fairness metrics, including Multi-group Equality of Opportunity (MEO), Disparate Impact (DI), and Maximum Disparity (MD) (section 3.3). For Hard ImageNet, which contains spurious correlations without known subgroup partitions, we measure model accuracy on images with various ablation masks applied to the spurious objects.

Implementations. We set the number of generations to 10 or 4 in MNIST/CIFAR and Hard ImageNet, respectively. For training all models, we use the Adam optimizer, initializing the learning rate at 1×10^{-1}, with training capped at 50 epochs. Early stopping is employed to ensure full convergence and to avoid overfitting. We provide the evaluation of different generators across generations based on the FID score in table 1. The classification model at the 0-the generation is trained on the original dataset without any generated data. The queue has a maximum capacity of 3. For all results, We run 3 times to reduce the experimental randomness.

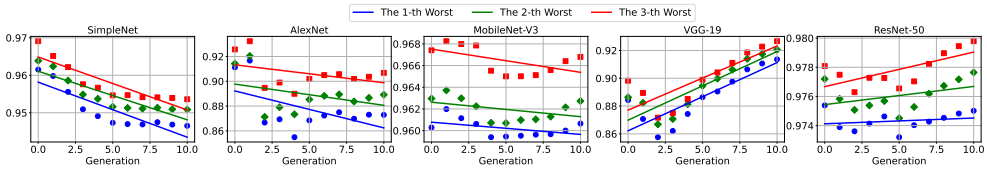
4.2 EVALUATION ON COLORIZED MNIST

We begin with the Colorized MNIST dataset, using both unbiased and biased initializations. The introduction of bias refers to the uneven painting strategy applied at the outset.

Unbiased Initialization. The results are shown in fig. 3 . Most models benefit from data augmentation using the updated generated data across generations, and all single-bias evaluations also show slight improvements. However, there are notable exceptions, particularly with MobileNet-V3, which experiences significant performance variations across generations. It's important to highlight that models differ considerably in multi-bias evaluations. While VGG-19 and ResNet-50 show significant improvements, smaller models, including SimpleNet, AlexNet, and MobileNet-V3, exhibit a noticeable decline in subgroup performance with continued large generations.

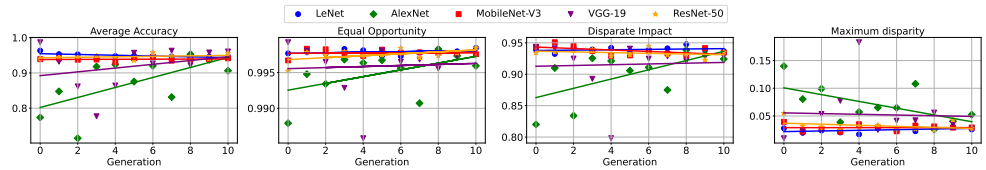


(a) Overall performance of the model trained on the Colorized MNIST dataset with unbiased initialization.

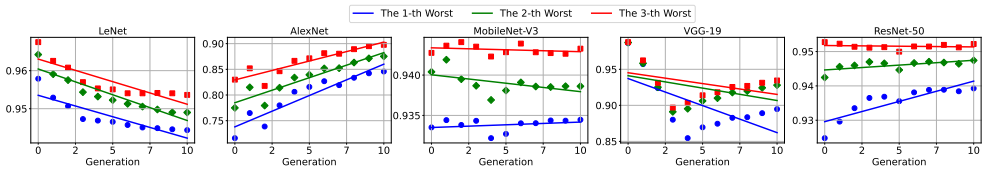


(b) Subgroup performance of the model trained on the Colorized MNIST dataset with unbiased initialization.

Figure 3: Results on the models trained on the MNIST dataset with unbiased initialization.



(a) Overall performance of the model trained on the Colorized MNIST dataset with biased initialization.

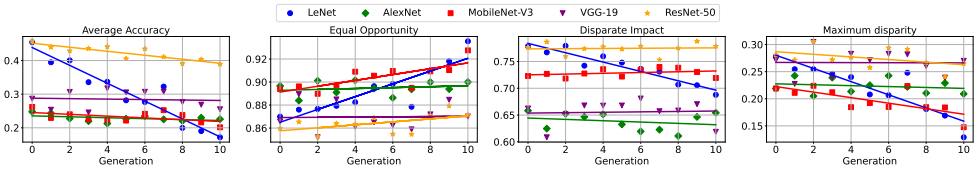


(b) Subgroup performance of the model trained on the Colorized MNIST dataset with biased initialization.

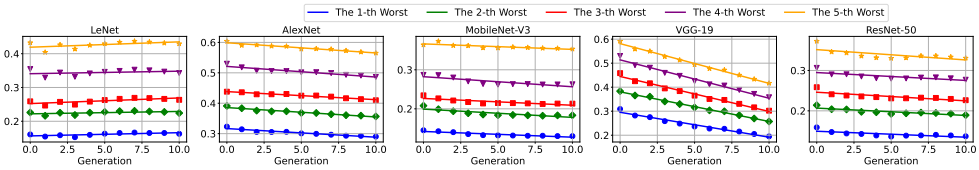
Figure 4: Results on the models trained on the MNIST dataset with biased initialization.

Biased Initialization. We present the results of models trained on the dataset with biased initialization in fig. 4. We observe consistent results in terms of classification accuracy and single-bias evaluation, which presents continuously imrprovement across generations; however, there are significant differences in the multi-bias evaluation. Specifically, VGG-19 experiences substantial performance degradation across subgroups, despite improvements on the dataset with unbiased initialization. In contrast, AlexNet performs better on this dataset as the number of generations increases. Compared with the results on the colorized MNIST with unbiased initialization, though MobileNet-V3 presents stable performance in this environment, both of the AlexNet and VGG-19 show large variation.

Summarization & Takeaways. As reported in table 1, the generative model can learn an approaching latent representation similar to that of the real samples on the colorized MNIST datasets, which is evident by the similar results on the FID evolution across generations. Thus, we can omit the impact of the quality of the generated data on the downstream models here. On the MNIST dataset, models can be consistently improved by augmenting the dataset with generated data across multiple generations. Notably, the inclusion of additional generated data does not significantly affect the models' single-bias performance, even with a large number of generations. However, it can lead to substantial variations in subgroup performance, revealing the presence of the multi-bias problem. The impact of generated data across generations varies between different models but remains consistent within the same architecture over multiple generations. Comparing results from unbiased and biased initializations, we observe that the presence of bias in the original dataset does not cause the model to degrade rapidly. Both initialization types exhibit similar trends in single- and multi-bias performance. In other words, the presence of dataset bias does not significantly amplify model bias when the dataset is augmented with generated data across generations.



(a) Overall performance of the model trained from scratch on the CIFAR-20/100 dataset.



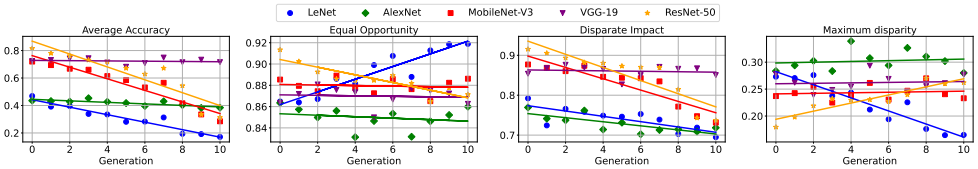
(b) Subgroup performance of the model trained from scratch on the CIFAR-20/100 dataset.

Figure 5: Results on the models trained from scratch on the CIFAR-20/100 dataset.

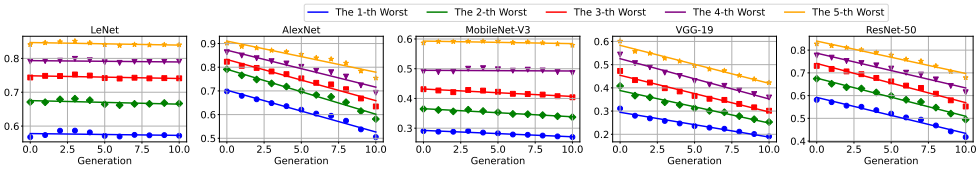
4.3 Evaluation on CIFAR-20/100

Next, we proceed to a more challenging dataset, CIFAR-20/100. Different from MNIST, the original CIFAR dataset comprises more features and biases influencing the model training, which are not easily controllable. Thus, we investigate the impact of using pre-trained weights on the model bias during the self-consuming loop. In this experiment, we compare the performance of models initialized with pre-trained weights provided by the PyTorch library to those trained from scratch. This comparison will help assess the effectiveness of pre-trained weights in improving model performance and stability when applied in this iterative data augmentation process.

Without Pre-trained Weights. Unlike the results on the MNIST dataset, augmenting CIFAR-20/100 with generated data can lead to degradation, with LeNet experiencing up to a 20% drop after 10 generations. The impact on bias metrics also varies. In the single-bias evaluation, both Equality of Opportunity and Maximum Disparity are significantly improved across all models, while most models show similar behavior regarding Disparate Impact. LeNet exhibits a larger bias in terms of Disparate Impact. For the multi-bias evaluation, models perform more consistently across different subgroups compared to their average performance across generations. Notably, although VGG-19 shows decreasing performance over generations, it performs better in bridging the performance gap between different subgroups.



(a) Overall performance evaluation of the fine-tuned model on the CIFAR-100.



(b) Subgroup performance evaluation of the fine-tuned model on the CIFAR-100 dataset.

Figure 6: Results on the models pre-trained on the ImageNet and fine-tuned on the CIFAR-20/100.

With Pre-trained Weights.Notably, models perform a faster performance degradation when using pre-trained weights as the number of generations for data augmentation increases. A greater number of models exhibit declines in classification accuracy and fairness metrics, such as Equality of Opportunity and Disparate Impact. Interestingly, while ResNet-50 without pretraining does not show a significant performance drop in the multi-bias evaluation, it experiences substantial degradation when pre-trained on the ImageNet dataset. This suggests that pre-trained weights, despite their ini-

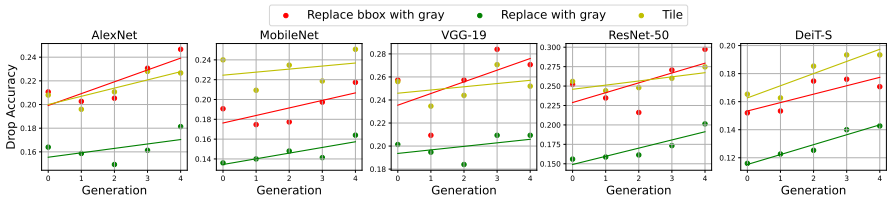


Figure 8: Evaluations of the impact of spurious correlation on the models pre-trained on the ImageNet and fine-tuned on the augmented Hard-ImageNet dataset across generations.

tial advantage, may exacerbate model bias and performance issues in this iterative augmentation process.

Summarization & Takeaways. As shown in table 1, continuously training on the dataset augmented by generated data across multiple generations leads to a slight improvement in generative performance, as evidenced by the decreasing FID scores on the CIFAR-20/100 dataset. However, despite the improved generative model, classification models trained with successively augmented datasets still experience a decline in performance in both the original classification task and bias evaluations. When using pre-trained weights from the ImageNet dataset, the classification models show significant improvement compared to training from scratch. Nevertheless, it is evident that models with pre-trained weights are more susceptible to integration bias introduced by the augmented datasets evolved over generations, further exacerbating performance deterioration in bias evaluations.

4.4 Evaluation on Hard ImageNet

We also conduct experiments on Hard Imagenet (Moayeri et al., 2022a), a dataset gathered from ImageNet with very strong spurious cues. The dataset contains 15 classes, and in each class, there is a strong correlation between the image background and the objects. This may lead the model to rely on background information rather than the actual objects for classification. Compared to the pre-defined color bias in Colorized MNIST and existing subgroup biases in the CIFAR-20/100 dataset, the unknown spurious correlation bias in this dataset is more challenging and difficult to fully identify, making it harder to mitigate during model development.

To study the impact of cross-generational data on this model bias, we made a modification to our proposed simulation framework. First, we fine-tuned the Stable Diffusion model using Low-Rank Adaptation rather than training from scratch to achieve a good balance between efficiency and generation quality on our task. Then we use 5 generations of mixed datasets to fine-tune our classifiers. Subsequently, while lacking explicit signal for single and multiple bias attributes, ablation studies are conducted on each classifier, following the approach described in Moayeri et al. (2022a). Specifically, we performed three types of ablation: (1) the object pixels were replaced with a uniform value of 0.5, neutralizing the object's appearance; (2) the entire bounding box surrounding the object was replaced with gray, removing shape-related information, and (3) the bounding box was replaced with a neighboring region of the image, substituting the object with local context. The performance drop caused by masking the image can indicate the model's reliance on spurious correlations. A significant performance drop suggests that the model's predictions rely more on the core object, indicating less influence from spurious correlations.

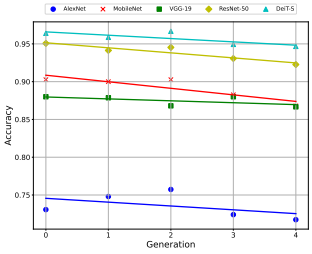


Figure 7: Subgroup accuracy of models fine-tuned on the augmented dataset across generations.

We report the changes in classification accuracy across generations in fig. 9 and the impact on learned spurious correlations in fig. 8. Over time, we observe that generated data can degrade model performance, as evidenced by the negative correlation between performance and the number of generations. However, for smaller models, performance sees a notable improvement during the first two generations, after which it stabilizes or becomes slightly better than the initial generation. Regarding bias evaluation, all models show a tendency to rely less on spurious correlations, indicating a shift toward focusing more on the core object for classification.

5 WHY MODELS EXHIBIT DIVERSE BEHAVIORS ACROSS GENERATIONS

The varied behaviors observed across different datasets and models can be attributed to several factors, including the datasets, models, and data quality across generations. These factors interact with each other in complex ways, influencing the dynamics of bias across generations.

Dataset Characteristics. Different datasets exhibit unique features such as image complexity, class diversity, and inherent biases. Let β_D represent the inherent bias in the dataset. For simpler datasets like Colorized MNIST, generative models can learn accurate representations more easily, resulting in generated data that closely matches the original data distribution. This closeness can be quantified by a high data quality factor $q_t \approx 1$ at generation t. The generated data with minimized bias helps the model continuously improve its classification performance and reduce bias.

Model Architecture Sensitivity. Different model architectures have varying capacities to learn from augmented data and mitigate bias. Let γ_M represent the model's capacity to mitigate bias, which is a function of the model's architecture M. Larger models with higher capacity (e.g., VGG-19, ResNet-50) have higher γ_M, enabling them to handle biases in the data better. Conversely, smaller models (e.g., LeNet, AlexNet) have lower γ_M and are more susceptible to biases in the training data, leading to greater performance variability across generations.

Exposure of Bias. Datasets contain various biases, both known and unknown, explicit or difficult to detect. The exposure of bias can be represented by a bias amplification factor δ, which accounts for the complexity and ingrained biases within the dataset. As biases become more difficult to identify—progressing from color bias to subgroup bias and spurious correlations—we observe greater fluctuations in model performance. The bias in the model at generation $t+1$, denoted $B_{\text{model}}^{(t+1)}$, can be influenced by the bias in the data and the model's capacity to mitigate it.

Unbalanced Generation. As identified in previous studies (Sehwag et al., 2022; Lee et al., 2021), generative models typically generate data from high-density regions of the data distribution, potentially over-representing certain classes or features. This tendency can be represented by an unbalanced generation factor u_t at generation t, which contributes to the bias in the generated data. The quality of data generation is crucial; lower-quality data can degrade the overall representation quality, which may mitigate biased performance in downstream models by introducing noise.

Combining these factors, we have a conjecture about modeling the bias dynamics across generations using a recursive relationship. Let the bias in the model at generation $t+1$ be expressed as:

$$B_{\text{model}}^{(t+1)} = (1-\gamma_M)\,(1+\delta_D+\delta_Q(1-q_t)+\delta_U u_t)\,B_{\text{model}}^{(t)}. \tag{6}$$

Then, the overall bias amplification factor A_t at the generation t can be denoted as $A_t := (1-\gamma_M)\,(1+\delta_D+\delta_Q(1-q_t)+\delta_U u_t)$. Depending on the values of γ_M, q_t, u_t, and the constants δ_D, δ_Q, and δ_U, the bias amplification factor A_t can be greater or less than 1. If $A_t > 1$, the bias increases across generations; if $A_t < 1$, the bias decreases.

Thus, the interplay between dataset characteristics, model architecture sensitivity, exposure of bias, and unbalanced generation may determine the bias dynamics across generations. To establish a self-sustaining model development loop with positive feedback, it is essential to have a clearer understanding of dataset bias (δ_D), utilize larger models with higher capacity (γ_M), and employ high-quality generative models with improved sampling mechanisms to increase q_t and reduce u_t.

6 CONCLUSION

Several models, like Stable Diffusion (Rombach et al., 2022), LLaMA (Touvron et al., 2023), LLaVA (Liu et al., 2024), and Nemotron (Adler et al., 2024), involve self-consumption loops. Notably, Nemotron is trained with over 98% synthetic data. While synthetic data can improve training, it may also introduce risks, particularly related to model biases. This has led us to investigate how generated data affects model performance and bias, especially as self-consumption loops increase. Our experiments on Colorized MNIST, CIFAR-20/100, and Hard ImageNet datasets show that bias changes depend on factors like dataset type, model architecture, and generative model performance. Additionally, models are more sensitive to multiple biases than to a single one.

REFERENCES

Bo Adler, Niket Agarwal, Ashwath Aithal, Dong H Anh, Pallab Bhattacharya, Annika Brundyn, Jared Casper, Bryan Catanzaro, Sharon Clay, Jonathan Cohen, et al. Nemotron-4 340b technical report. *arXiv preprint arXiv:2406.11704*, 2024.

Yaganteeswarudu Akkem, Saroj Kumar Biswas, and Aruna Varanasi. A comprehensive review of synthetic data generation in smart farming by using variational autoencoder and generative adversarial network. *Engineering Applications of Artificial Intelligence*, 131:107881, 2024.

Sina Alemohammad, Josue Casco-Rodriguez, Lorenzo Luzi, Ahmed Imtiaz Humayun, Hossein Babaei, Daniel LeJeune, Ali Siahkoohi, and Richard Baraniuk. Self-consuming generative models go mad. In *The Twelfth International Conference on Learning Representations*, 2024.

Krizhevsky Alex. Learning multiple layers of features from tiny images. *https://www. cs. toronto. edu/kriz/learning-features-2009-TR. pdf*, 2009.

Laith Alzubaidi, Jinshuai Bai, Aiman Al-Sabaawi, Jose Santamaría, Ahmed Shihab Albahri, Bashar Sami Nayyef Al-dabbagh, Mohammed A Fadhel, Mohamed Manoufali, Jinglan Zhang, Ali H Al-Timemy, et al. A survey on deep learning tools dealing with data scarcity: definitions, challenges, solutions, tips, and applications. *Journal of Big Data*, 10(1):46, 2023.

Antreas Antoniou, Amos Storkey, and Harrison Edwards. Data augmentation generative adversarial networks. *arXiv preprint arXiv:1711.04340*, 2017.

Shekoofeh Azizi, Simon Kornblith, Chitwan Saharia, Mohammad Norouzi, and David J Fleet. Synthetic data from diffusion models improves imagenet classification. *Transactions on Machine Learning Research*, 2023.

Dariusz Brzezinski, Julia Stachowiak, Jerzy Stefanowski, Izabela Szczech, Robert Susmaga, Sofya Aksenyuk, Uladzimir Ivashka, and Oleksandr Yasinskyi. Properties of fairness measures in the context of varying class imbalance and protected group ratios. *ACM Transactions on Knowledge Discovery from Data*, 2024.

Zixiang Chen, Yihe Deng, Huizhuo Yuan, Kaixuan Ji, and Quanquan Gu. Self-play fine-tuning converts weak language models to strong language models. *arXiv preprint arXiv:2401.01335*, 2024.

Florinel-Alin Croitoru, Vlad Hondru, Radu Tudor Ionescu, and Mubarak Shah. Diffusion models in vision: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 45(9): 10850–10869, 2023.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *2009 IEEE conference on computer vision and pattern recognition*, pp. 248–255. Ieee, 2009.

Xiang Deng, Yu Gu, Boyuan Zheng, Shijie Chen, Sam Stevens, Boshi Wang, Huan Sun, and Yu Su. Mind2web: Towards a generalist agent for the web. *Advances in Neural Information Processing Systems*, 36, 2024.

Lijie Fan, Kaifeng Chen, Dilip Krishnan, Dina Katabi, Phillip Isola, and Yonglong Tian. Scaling laws of synthetic images for model training... for now. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 7382–7392, 2024.

Michael Feldman, Sorelle A Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. Certifying and removing disparate impact. In *proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining*, pp. 259–268, 2015.

Francisco HG Ferreira and Vito Peragine. Equality of opportunity: Theory and evidence. 2013.

Sorelle A Friedler, Carlos Scheidegger, Suresh Venkatasubramanian, Sonam Choudhary, Evan P Hamilton, and Derek Roth. A comparative study of fairness-enhancing interventions in machine learning. In *Proceedings of the conference on fairness, accountability, and transparency*, pp. 329–338, 2019.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *international conference on machine learning*, pp. 1050–1059. PMLR, 2016.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

David Gray Grant. Equalized odds is a requirement of algorithmic fairness. *Synthese*, 201(3):101, 2023.

Hasan Abed Al Kader Hammoud, Hani Itani, Fabio Pizzati, Adel Bibi, and Bernard Ghanem. Synthclip: Are we ready for a fully synthetic clip training? In *Synthetic Data for Computer Vision Workshop@ CVPR 2024*.

Chunming He, Yuqi Shen, Chengyu Fang, Fengyang Xiao, Longxiang Tang, Yulun Zhang, Wangmeng Zuo, Zhenhua Guo, and Xiu Li. Diffusion models in low-level vision: A survey. *arXiv preprint arXiv:2406.11138*, 2024.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 770–778, 2016.

Edgar Heinert, Matthias Rottmann, Kira Maag, and Karsten Kahl. Reducing texture bias of deep neural networks via edge enhancing diffusion. *arXiv preprint arXiv:2402.09530*, 2024.

Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. Gans trained by a two time-scale update rule converge to a local nash equilibrium. *Advances in neural information processing systems*, 30, 2017.

Peter Hönig, Stefan Thalhammer, Jean-Baptiste Weibel, Matthias Hirschmanner, and Markus Vincze. Star: Shape-focused texture agnostic representations for improved object detection and 6d pose estimation. *arXiv preprint arXiv:2402.04878*, 2024.

Hossein Hosseini, Baicen Xiao, Mayoore Jaiswal, and Radha Poovendran. Assessing shape bias property of convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pp. 1923–1931, 2018.

Andrew Howard, Mark Sandler, Grace Chu, Liang-Chieh Chen, Bo Chen, Mingxing Tan, Weijun Wang, Yukun Zhu, Ruoming Pang, Vijay Vasudevan, et al. Searching for mobilenetv3. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 1314–1324, 2019.

Zhimeng Jiang, Xiaotian Han, Chao Fan, Fan Yang, Ali Mostafavi, and Xia Hu. Generalized demographic parity for group fairness. In *International Conference on Learning Representations*, 2022.

Sangwon Jung, Donggyu Lee, Taeeon Park, and Taesup Moon. Fair feature distillation for visual recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 12115–12124, 2021.

Mahyar Khayatkhoei and Ahmed Elgammal. Spatial frequency bias in convolutional generative adversarial networks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 7152–7159, 2022.

Byungju Kim, Hyunwoo Kim, Kyungsu Kim, Sungjin Kim, and Junmo Kim. Learning not to learn: Training deep neural networks with biased data. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 9012–9020, 2019a.

Michael P Kim, Amirata Ghorbani, and James Zou. Multiaccuracy: Black-box post-processing for fairness in classification. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 247–254, 2019b.

Hadas Kotek, Rikker Dockum, and David Sun. Gender bias and stereotypes in large language models. In *Proceedings of The ACM Collective Intelligence Conference*, pp. 12–24, 2023.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

Yann LeCun. The mnist database of handwritten digits. *http://yann. lecun. com/exdb/mnist/*, 1998.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

Jinhee Lee, Haeri Kim, Youngkyu Hong, and Hye Won Chung. Self-diagnosing gan: Diagnosing underrepresented samples in generative adversarial networks. *Advances in Neural Information Processing Systems*, 34:1925–1938, 2021.

Yang Li, Quan Pan, Suhang Wang, Tao Yang, and Erik Cambria. A generative model for category text generation. *Information Sciences*, 450:301–315, 2018.

Yiyi Liao, Katja Schwarz, Lars Mescheder, and Andreas Geiger. Towards unsupervised learning of generative models for 3d controllable image synthesis. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 5871–5880, 2020.

Xiaofeng Lin, Seungbae Kim, and Jungseock Joo. Fairgrape: Fairness-aware gradient pruning method for face attribute classification. In *European Conference on Computer Vision*, pp. 414–432. Springer, 2022.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36, 2024.

Ruibo Liu, Chenyan Jia, Jason Wei, Guangxuan Xu, and Soroush Vosoughi. Quantifying and alleviating political bias in language models. *Artificial Intelligence*, 304:103654, 2022.

Orestis Loukas and Ho-Ryun Chung. Demographic parity: Mitigating biases in real-world data. *arXiv preprint arXiv:2309.17347*, 2023.

Cong Lu, Philip Ball, Yee Whye Teh, and Jack Parker-Holder. Synthetic experience replay. *Advances in Neural Information Processing Systems*, 36, 2024.

Carmen Mazijn, Jan Danckaert, and Vincent Ginis. How do the score distributions of subpopulations influence fairness notions? In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 767–776, 2021.

Yu Meng, Jiaxin Huang, Yu Zhang, and Jiawei Han. Generating training data with language models: Towards zero-shot language understanding. *Advances in Neural Information Processing Systems*, 35:462–477, 2022.

Mazda Moayeri, Sahil Singla, and Soheil Feizi. Hard imagenet: Segmentations for objects with strong spurious cues. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 10068–10077. Curran Associates, Inc., 2022a. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/4120362f930b0b683cd30b71af56fad1-Paper-Datasets_and_Benchmarks.pdf.

Mazda Moayeri, Sahil Singla, and Soheil Feizi. Hard imagenet: Segmentations for objects with strong spurious cues. *Advances in Neural Information Processing Systems*, 35:10068–10077, 2022b.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

Alec Radford. Unsupervised representation learning with deep convolutional generative adversarial networks. *arXiv preprint arXiv:1511.06434*, 2015.

Yuji Roh, Kangwook Lee, Steven Euijong Whang, and Changho Suh. Fairbatch: Batch selection for model fairness. *arXiv preprint arXiv:2012.01696*, 2020.

Yaniv Romano, Stephen Bates, and Emmanuel Candes. Achieving equalized odds by resampling sensitive attributes. *Advances in neural information processing systems*, 33:361–371, 2020.

Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pp. 10684–10695, 2022.

Vikash Sehwag, Caner Hazirbas, Albert Gordo, Firat Ozgenel, and Cristian Canton. Generating high fidelity data from low-density regions using diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 11492–11501, 2022.

Seonguk Seo, Joon-Young Lee, and Bohyung Han. Information-theoretic bias reduction via causal view of spurious correlation. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 36, pp. 2180–2188, 2022.

Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

Krishnakant Singh, Thanush Navaratnam, Jannik Holmer, Simone Schaub-Meyer, and Stefan Roth. Is synthetic data all we need? benchmarking the robustness of models trained with synthetic images. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 2505–2515, 2024.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. Going deeper with convolutions. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 1–9, 2015.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In *International conference on machine learning*, pp. 10347–10357. PMLR, 2021.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Yixin Wang, Dhanya Sridhar, and David Blei. Adjusting machine learning decisions for equal opportunity and counterfactual fairness. *Transactions on Machine Learning Research*, 2023.

Zuhao Yang, Fangneng Zhan, Kunhao Liu, Muyu Xu, and Shijian Lu. Ai-generated images as data source: The dawn of synthetic era. *arXiv preprint arXiv:2310.01830*, 2023.

Zeliang Zhang, Mingqian Feng, Zhiheng Li, and Chenliang Xu. Discover and mitigate multiple biased subgroups in image classifiers. *arXiv preprint arXiv:2403.12777*, 2024.

Chenyu Zheng, Guoqiang Wu, and Chongxuan Li. Toward understanding generative data augmentation. *Advances in Neural Information Processing Systems*, 36, 2024.

Yongchao Zhou, Hshmat Sahak, and Jimmy Ba. Training on thin air: Improve image classification with generated data. *arXiv preprint arXiv:2305.15316*, 2023.

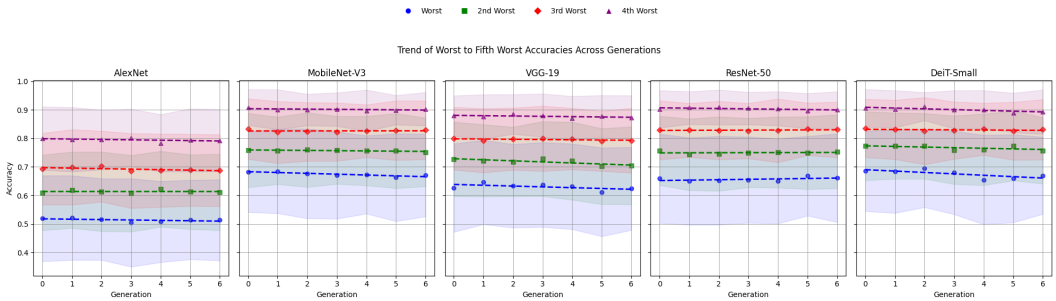


Figure 9: Classification accuracy of models fine-tuned on the augmented ImageNet across generations.

A More results on the ImageNet

We conduct additional experiments on the ImageNet dataset Breeds, which is organized by subclasses as defined by WordNet. Each superclass consists of four subclasses. Following the same settings used for Hard-ImageNet, we utilize Stable Diffusion 1.5 to learn the dataset's distribution and augment it with generated data across multiple generations. In each generation, we train AlexNet, MobileNet-V3, VGG-19, ResNet-50, and DeiT-Small. The results are shown in fig. 7.

We observe a consistent phenomenon with the results on Hard-ImageNet. Compared to the best-performing subgroup, the generated data has a greater impact on the worst-performing subgroups, as indicated by a steeper slope across different generations.

B Examples of generated images across generations

As shown in fig. 10, fig. 11, and fig. 12, each row represents a generation of images, with the generation number increasing sequentially from top to bottom. We can find that on MNIST and CIFAR-20/100 dataset, the quality of generated data doesn't change a lot, while it decreases significantly for the Hard ImageNet dataset.

C Details on the expert-guided filtering

First, we manually review the generated samples and discard images with low quality.

Second, we calculate the CLIP score for each image, where the paired text is the class name. Images are then grouped into bins based on their CLIP scores, with each bin representing a $\pm 10\%$ range of CLIP scores. This results in 10 bins.

Then, we randomly sample 10 images from each bin and evaluate the quality of each bin. Based on this evaluation, we determine the maximum ratio of the CLIP score range (denoted as $r\%$) to retain for training.

- For MNIST, we find that retaining the top 90% of images ($r = 10\%$) is optimal.
- For CIFAR-20/100, retaining the top 70% of images ($r = 30\%$) works best.
- For the ImageNet dataset, retaining the top 40% of images ($r = 60\%$) yields the best results.

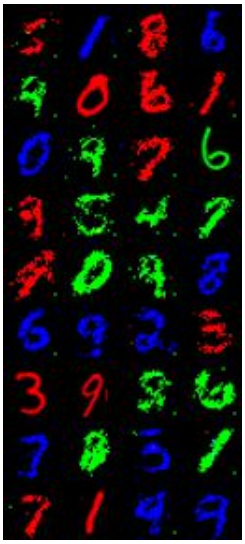


Figure 10: Color-MNIST

Figure 11: CIFAR-20/100

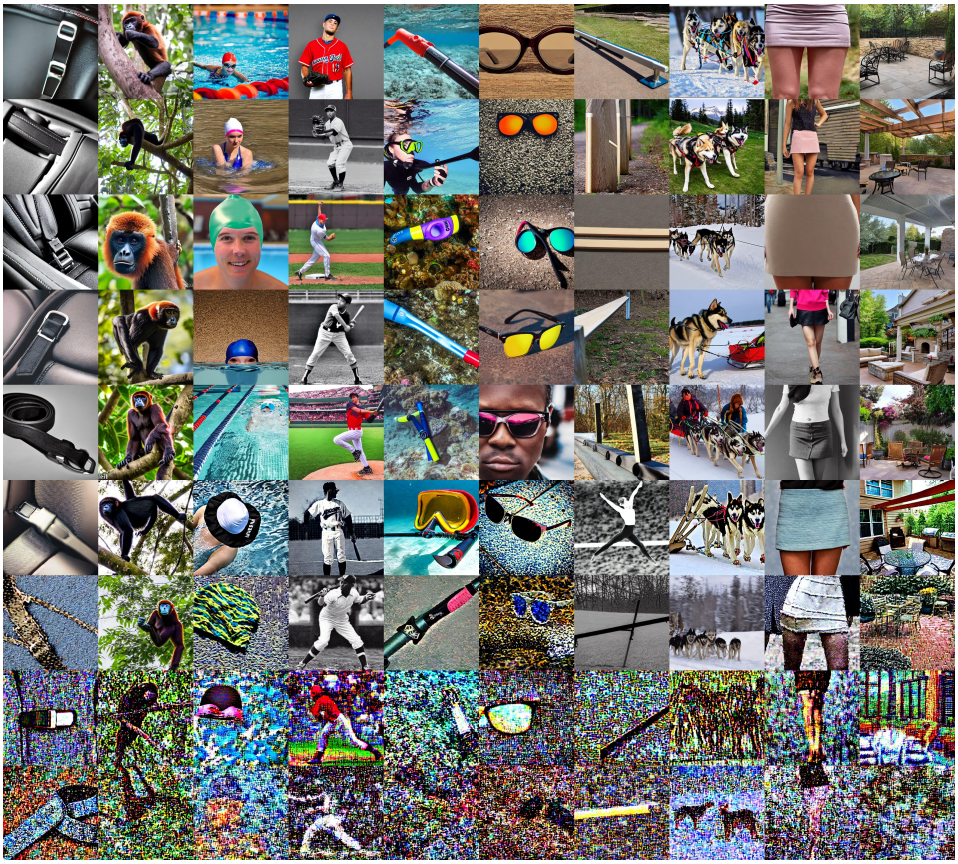


Figure 12: Hard ImageNet