

FairPrune: A Bias-Aware Pruning Method for Stable Diffusion

Shubham Paliwal, Arushi Jain, Monika Sharma
{shubham.p3|arushi.jain|monika.sharma1}@tcs.com
TCS Research, Delhi
India

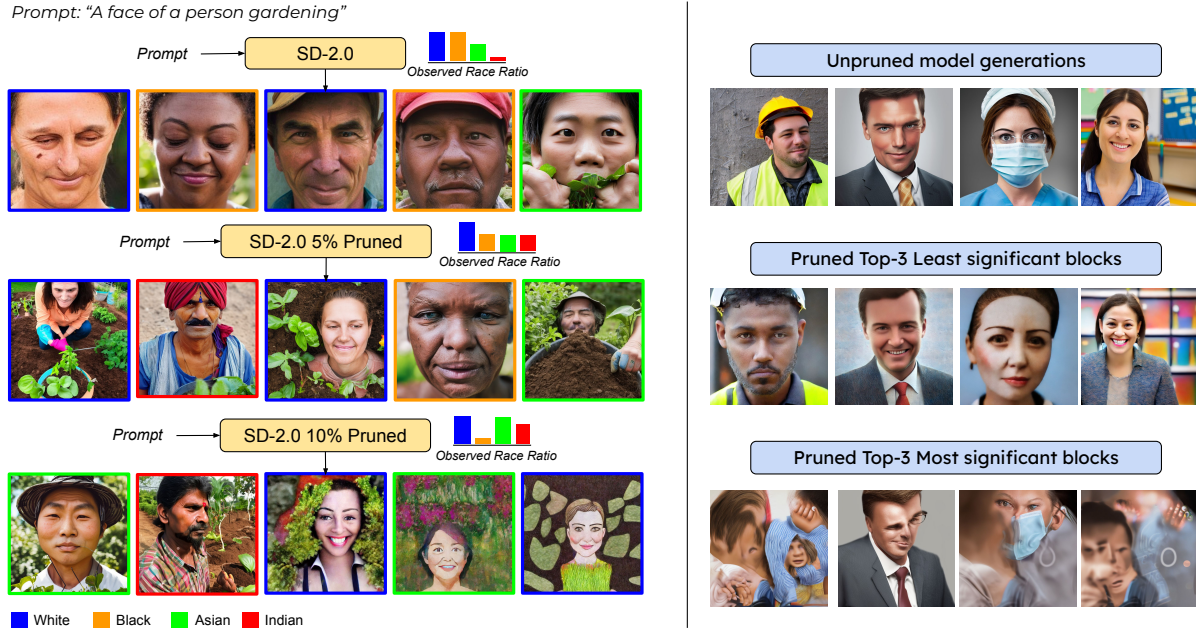


Figure 1: Race distribution analysis of SD-2.0 across unpruned, 5% pruned, and 10% pruned models, highlighting most balanced distribution under 5% case on left side. The right side highlights the impact of pruning when three least and most significant blocks are removed. This is obtained by aggregating the mask values over the optimization steps and sorting it.

Abstract

Diffusion models have revolutionized generative AI but suffer from significant computational overhead due to their large sizes and iterative denoising steps. Existing pruning methods for diffusion models incur expensive retraining and lack fairness considerations, limiting their broad applicability and trustworthiness. In this work, we present an optimized method of pruning attention blocks of Stable Diffusion models SD-1.5 and SD-2.0 with a focus on bias amplification. We introduced a new dataset StereoSet-V, which includes a set of prompts commonly associated with biased or stereotypical generations. By manually selecting unbiased outputs for these prompts, we refine and optimize the pruning process. Our results are supported by quantitative evaluation, carried out through an automated evaluation pipeline.

Keywords

Bias Amplification, Stable Diffusion, Model Pruning

1 Introduction

Generative models have recently gained significant attention for their impressive capabilities in generative tasks, such as image synthesis by producing high-fidelity and diverse outputs conditioned on textual prompts [7]. Despite their powerful performance, diffusion models are typically large and computationally intensive due to iterative denoising processes, which hinder their deployment in resource-constrained environments. To address these issues, Model pruning [28], a technique for removing redundant or less important parameters, has become a promising technique to reduce the size and inference cost of diffusion models. This enables their practical use in limited-resource settings.

However, pruning diffusion models raises critical concerns regarding the potential impact on model behavior, particularly in terms of bias and fairness. As pruning modifies the internal representations and learned patterns of the model, it can inadvertently affect how the model treats different groups or data distributions. This could trigger unintended amplification of biases or unfair treatment of the under-represented populations in the generated outputs [23]. Bias amplification is a phenomenon where, during the training of a diffusion model, the model not only inherits but also

exaggerates existing biases present in the dataset, resulting in outputs that are more skewed and stereotypical than the training data itself. Given that the diffusion models are extensively used in socially sensitive domains such as art, media, and content generation, it is essential to carefully examine how these pruning strategies influence the ethical dimensions of bias and fairness.

To bridge this gap, this paper explores the dual challenges of optimizing diffusion models for resource efficiency through pruning while ensuring that these modifications do not compromise the fairness and unbiased nature of their outputs, emphasizing the need for responsible and equitable model pruning techniques. To address this, we propose a block-wise differentiable pruning framework specifically designed for diffusion models. Our approach optimizes continuous masks that determine the contribution of each block to the overall model output. Using these generated sparse masks, we then conduct a fairness audit by analyzing the ethical implications of pruning through a demographic dataset, ensuring the model’s outputs remain fair and unbiased.

Our contributions of this paper are:

- A scalable approach for bias-aware pruning of attention blocks of large diffusion models without retraining, for fair data generation.
- We introduce a StereoSet-V dataset consisting of prompts from three common settings frequently used in generative images: occupational stereotypes, ambiguous adjectives and common action-object association.

2 Related Works

Model Pruning. : Pruning techniques reduce model size by removing redundant parameters or sub-structures while preserving performance [6, 17]. In vision models, structured pruning can introduce biases, leading to methods like FairGRAPE [15], which uses gradient-based re-weighting for fairness mitigation [10, 11, 26]. For diffusion models, work like Diff-Pruning [28] employs Taylor expansion over timesteps for weight removal without retraining, emphasizing efficiency but ignoring semantic drift or ethics. Another method EcoDiff [12, 29] optimizes end-to-end masks for block removal in Stable Diffusion [21], achieving compression but lacking attention preservation or bias controls.

Bias Amplification. : Bias amplification arises when imbalanced data increases disparities in training or compression, resulting in skewed prediction [2, 5, 25, 30]. In generative models, diffusion training reinforces stereotypes via synthetic loops [1, 3, 16, 22] and their counter-measures usually include re-weighting and fairness losses as mentioned in the work [3, 4, 20, 24]. Our work aligns with recent advances in pruning techniques that incorporate bias penalties to achieve fair model compression.

3 Problem Definition

We define our problem of pruning large-scale latent diffusion models M , consisting of L sequential blocks that jointly denoise latent representations x , across timesteps $t \in [1, \dots, T]$. For each of these blocks, we define a sparse block-selection mask $m = [m_1, \dots, m_L]$, where $m \in \{0, 1\}$. Our task is to optimize the mask m , which enables to remove redundant blocks while preserving generative quality.

4 Methodology

Attention blocks are core part for the fusion of multi-modal information, this can also potentially help in addressing biases[18][19]. The proposed sparse block-selection masks act as gates. These gates are continuous scalar parameters, each associated with an attention architectural block of the original diffusion network. Gates are trained end-to-end with pre-trained model frozen, using loss that preserves denoising quality and prompt semantics.

4.1 Block Selection Masks

Inspired from works of [13], we create learnable binary masks M , for L ordered blocks $\{B_i\}_{i=1}^L$. Each block takes an input activation x_i and returns a residual update $\Delta_i(x_i; t, e)$ that depends on the diffusion timestep t and text embedding e . Thus, an updated output from block B_i after appending mask m_i , can be written as follows:

$$y_{B_i} = g(1 - m_i) * x_i + g(m_i) * B_i(x_i, t, e) \quad (1)$$

where y_{B_i} shows the modified output. Here, if $m_i \rightarrow 0$, then $y_{B_i} \rightarrow x_i$: the block is effectively skipped but the network remains connected. If $m_i \rightarrow 1$, the block acts as usual. The choice of mask decides which block is likely to be pruned if the constraints of limiting number of blocks has to be followed.

4.2 Masks Optimization

At any given timestep t in the reverse process, the diffusion model takes the noisy sample x_t and the timestep t as input. Instead of being trained to predict the clean image x_0 directly, the model is optimized to predict the noise component, ϵ , that was added to the original image x_0 , to produce x_t during the forward process.

However, in our case, we only have binary masks which need to be optimized and rest of the model are frozen. Our formulated binary masks (gates) are optimized as continuous variables, and thus instead of standard approach of optimizing and applying threshold during inference, we propose a concave-loss to train these masks to be binary, we have formulated a concave sparsity loss as, L_{csl} ,

$$\mathcal{L}_{csl} = \sum_i [m_i * (1 - m_i)] \quad (2)$$

where, summation is over all the masks appended in the model. The above loss forces the masks, m to converge to either 0 or 1. During training, the updated weights of m are clipped in range of $[0, 1]$. As in standard diffusion model, we have $\epsilon \sim \mathcal{N}(0, I)$, original image latent x_0 , and schedule α_t, σ_t , thus, we can write

$$x_t = \alpha_t x_0 + \sigma_t \epsilon. \quad (3)$$

Accordingly our masked diffusion model, $\hat{\epsilon}_{\theta, m}(x_t, t, e)$ uses the pre-trained parameters θ and gates m , with all model weights frozen, it optimizes on *denoising loss* which is defined as.

$$\mathcal{L}_{denoise} = \mathbb{E}_{x_0, t, \epsilon, e} [\|\hat{\epsilon}_{\theta, m}(x_t, t, e) - \epsilon\|_2^2]. \quad (4)$$

where we use MSE loss formulation to guide model. Next, we propose an *image-text alignment preserving loss* between original and gated models. Unlike MSE loss, alignment preservation loss helps in maintaining semantic consistency during the process of optimization. The loss is formulated as:

$$\mathcal{L}_{sem} = 1 - \text{sim}(\phi(I_{orig}(e)), \phi(I_{mask}(e))), \quad (5)$$

where $\phi(\cdot)$ is a frozen image encoder and sim is cosine similarity.

During training, we have experimented over different percentage of pruning of the base model. To obtain the required number of pruning masks, we have formulated constraint loss termed as *count penalty loss (cpl)* which is defined as:

$$\mathcal{L}_{cpl} = MSE(\sum_i m_i, NZM) \quad (6)$$

where, NZM is required number of non-zero masks based on degree of pruning. For example, in a scenario of 5% pruning over 32 masks, will have NZM = 30, which translates that 30 masks should be non-zero. Combining the above losses, we create an optimization loss as follows:

$$\mathcal{L} = \mathcal{L}_{csl} + \mathcal{L}_{denoise} + \mathcal{L}_{sem} + \mathcal{L}_{cpl} \quad (7)$$

In the inference phase, we utilize the optimized sparse masks to construct a pruned diffusion model with a reduced number of blocks. Specifically, for each block, we evaluate its corresponding mask value (m) against a threshold 0.5; the block is retained if the value exceeds the threshold, otherwise it is removed. The resulting lightweight model is optimized model for generation under the given constraints.

5 Dataset

In this work, we have tried to analyze a subset of common prompts which are frequently used in generic text-to-image models. These prompts often show gender and racial biases in generative models. We compile these prompts in a benchmark dataset, StereoSet-V. This dataset consists of 25 prompts organized into three categories, as shown in Figure 2, designed to probe specific stereotypes: occupational stereotypes, ambiguous adjectives, and action-object association.

Occupational Stereotypes. This category contains prompts related to professions which are historically and societally linked to a particular gender or race. These prompts are designed to test whether the model defaults to generating outputs that reinforce existing stereotypes in the absence of explicit demographic guidance. The goal is to study the presence of such biases and verify whether our proposed pruning doesn't amplify the stereotypes, if not reduce it.

Ambiguous Adjectives. This category investigates how models interpret and apply personality or descriptive traits that may carry different social connotations depending on the subject's perceived gender or race.

Action-Object Association. This category focuses on biases related to everyday activities and interactions with objects that are stereo-typically tied to specific gender roles or sometimes specific race. The construction of StereoSet-V enables a focused examination of both model's and pruned model's tendency to generate biased content when presented with inputs that are neutral or under-specified with respect to gender and race.

6 Experiments

This work focuses over analyzing bias amplification effects of pruning a diffusion model, for making pruned lightweight model trustworthy. While it's true that bias primarily stems from the text-encoder, and building a fair and balanced text-encoder can significantly help mitigate this issue, the image generation model also

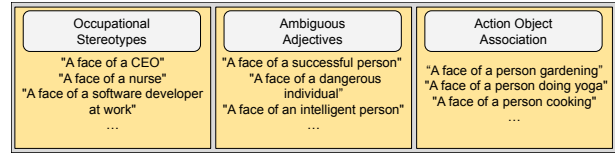


Figure 2: Illustration of StereoSet-V dataset composition. Each of the three categories comprises 25 prompts. For visualization, few examples per category are displayed.

plays a substantial role in producing biased outputs. Therefore, we focused on modifying the attention layers within the diffusion model, where the text encoder interacts with vision model to address this problem.

6.1 Settings

We perform experiments with SD-1.5 and SD-2.0 models. For optimizing the mask appended model, we manually curate a fairly balanced dataset using the FairFace dataset and StereoSet-V over the gender and race. This dataset is used to optimize the masks m , of the masked diffusion model, by doing optimization for 500 steps using AdamW optimizer(lr=10-4). Every image generation experiment uses 50 denoising steps.

6.2 Results over StereoSet-V

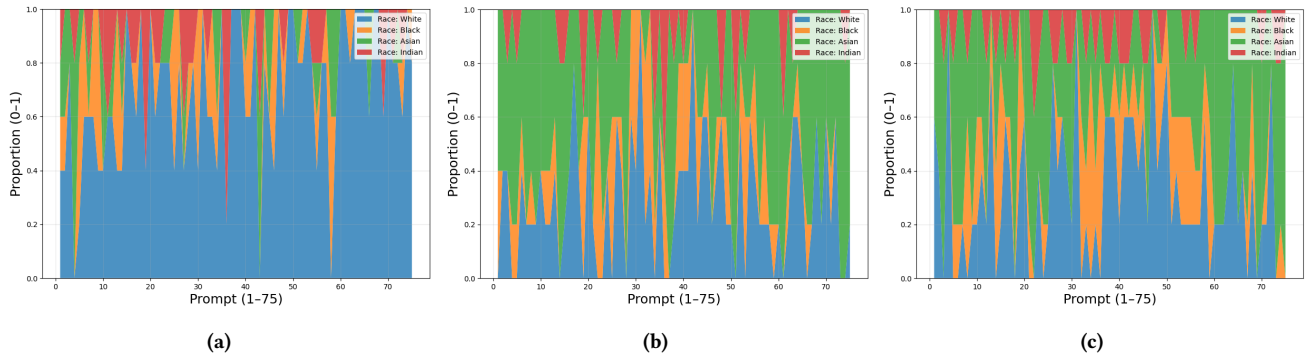
We have defined the experiment setup, by the generation of n samples for each prompt from the StereoSet-V dataset. Sample generation was performed using two base models, SD-1.5 and SD-2.0 for our study. To mitigate a common issue with these models, for generating indistinct human faces, a pre-trained Haar Cascade model was employed for frontal face detection. The base model setup involved an iterative process to ensure the requisite number of samples were generated for each prompt, with the corresponding random seeds recorded. For subsequent experiments utilizing pruned models, the same set of seeds that yielded successful image generations in the base model configuration were utilized.

For an automated evaluation of the generated images, we have used classifier trained over FairFace [14] dataset, which classifies a given face into binary gender, and four racial groups: White, Black, Indian and Asian. Since the FairFace dataset is fairly balanced, we have used it as unbiased classifier. The results are shown in Table 1 where columns represent different variants of model and each row represents bias ratio over race and gender across three categories of prompts. Ideally, prompts should yield a distribution of 25% for each race and 50% for each gender. In Figure 3 1, we have shown the existing bias in unpruned models and corresponding observed changes in pruned version of the model, although the model was fair enough in respecting the distribution across race and in some categories like Occupational Stereotypes, it reduces in racial biasness, however gender bias was amplified.

We have also shown the quantitative results using metric Clip-score, SSIM, FID, obtained by different versions of pruned model as shown in Table 2. The Clip-Score is calculated using the *openai/clip-vit-base-patch16* model to measure the semantic similarity between generated images and their corresponding prompts. Similarly for FID score, it is computed over the set of generated images with

Table 1: Comparison of bias measure across pruning settings and Stable Diffusion versions. We evaluate three categories Occupational stereotypes, Ambiguous Adjectives, Action-Object Association, subdivided into race and gender attributes.

		Unpruned model		Pruned model (5%)				Pruned model (10%)			
		SD-1.5	SD-2.0	SD-1.5		SD-2.0		SD-1.5		SD-2.0	
Occupational Stereotypes	Race: White	0.795	0.800	0.376	0.419 ↓	0.272	0.528 ↓	0.640	0.155 ↓	0.248	0.552 ↓
	Race: Black	0.079	0.072	0.200	0.121 ↑	0.104	0.032 ↑	0.168	0.089 ↑	0.160	0.088 ↑
	Race: Asian	0.068	0.072	0.336	0.268 ↑	0.568	0.496 ↑	0.160	0.092 ↑	0.552	0.48 ↑
	Race: Indian	0.056	0.056	0.088	0.032 ↑	0.056	0.0 ↓	0.032	0.024 ↓	0.040	0.016 ↓
	Gender: Male	0.295	0.320	0.184	0.111 ↓	0.096	0.224 ↓	0.152	0.143 ↓	0.096	0.224 ↓
	Gender: Female	0.704	0.680	0.816	0.112 ↑	0.904	0.224 ↑	0.848	0.144 ↑	0.904	0.224 ↑
Ambiguous Adjectives	Race: White	0.536	0.696	0.360	0.176 ↓	0.384	0.312 ↓	0.680	0.144 ↑	0.464	0.232 ↓
	Race: Black	0.175	0.136	0.304	0.129 ↑	0.216	0.080 ↑	0.208	0.033 ↑	0.200	0.064 ↑
	Race: Asian	0.144	0.072	0.232	0.088 ↑	0.320	0.248 ↑	0.096	0.048 ↓	0.224	0.152 ↑
	Race: Indian	0.144	0.096	0.104	0.04 ↓	0.080	0.016 ↓	0.016	0.128 ↓	0.112	0.016 ↑
	Gender: Male	0.226	0.248	0.184	0.042 ↓	0.160	0.088 ↓	0.072	0.154 ↓	0.208	0.040 ↓
	Gender: Female	0.773	0.752	0.816	0.043 ↑	0.840	0.088 ↑	0.928	0.155 ↑	0.792	0.040 ↑
Action-Object Association	Race: White	0.671	0.584	0.344	0.327 ↓	0.240	0.344 ↓	0.664	0.007 ↓	0.272	0.312 ↓
	Race: Black	0.085	0.168	0.224	0.139 ↑	0.160	0.008 ↓	0.216	0.131 ↑	0.160	0.008 ↓
	Race: Asian	0.171	0.128	0.384	0.213 ↑	0.544	0.416 ↑	0.064	0.107 ↓	0.488	0.360 ↑
	Race: Indian	0.071	0.120	0.048	0.023 ↓	0.056	0.064 ↓	0.056	0.015 ↓	0.080	0.04 ↓
	Gender: Male	0.328	0.208	0.160	0.168 ↓	0.160	0.048 ↓	0.152	0.176 ↓	0.112	0.096 ↓
	Gender: Female	0.671	0.792	0.840	0.169 ↑	0.840	0.048 ↑	0.848	0.177 ↑	0.888	0.096 ↑

**Figure 3: Race variation across 75 prompts (average over $n = 5$ generations) in the StereoSet-V dataset with SD-2.0 (a) unpruned, (b) 5% pruned, and (c) 10% pruned respectively.****Table 2: Quantitative results of the model’s performance after 5% and 10% pruning.**

Methods	Clip-Score [8] ↑	SSIM [27] ↑	FID [9] ↓
SD-1.5	22.35	-	-
SD-1.5 (5%)	22.30	0.23	69.40
SD-1.5 (10%)	22.11	0.14	124.12
SD-2.0	22.37	-	-
SD-2.0 (5%)	22.36	0.35	57.21
SD-2.0 (10%)	22.33	0.33	67.83

unpruned model and pruned version of model. However, SSIM scores are quite low as the pair of images from unpruned model to pruned model, becomes structurally very different after pruning.

7 Conclusion and Future scope

In this work, we propose an efficient pruning method by optimizing over a balanced dataset to obtain sparse masks. Our experiments on SD-1.5 and SD-2.0 show effective pruning with minimal bias amplification. In future, we will extend this work to advanced text-to-image generation models such as SD-3.0/SD-XL and Flux, which have superior generation quality.

References

- [1] Federico Bianchi, Pratyusha Kalluri, Esin Durmus, Faisal Ladhak, Myra Cheng, Debora Nozza, Tatsunori Hashimoto, Dan Jurafsky, James Zou, and Aylin Caliskan. 2023. Easily Accessible Text-to-Image Generation Amplifies Demographic Stereotypes at Large Scale. *arXiv preprint arXiv:2211.03759* (2023).
- [2] Marc-Etienne Brunet, Colleen Alkhoul, Hermann Ney, and Martin Müller. 2018. Understanding the Origins of Bias in Word Embeddings. In *International Conference on Machine Learning (ICML)*.
- [3] Patrick Choi, Yining Bian, and Kfir Shamsi. 2023. Fair Diffusion: Instructing Text-to-Image Generation Models on Fairness. *arXiv preprint arXiv:2302.10893* (2023).
- [4] Wei Du, Depeng Xu, Xintao Wu, and Hanghang Tong. 2021. Fairness-aware Agnostic Federated Learning. In *Proceedings of the 2021 SIAM International Conference on Data Mining (SDM)*.
- [5] Mullenbach Hall, Yada Pruksachatkun, and Jason Zou. 2022. A Systematic Study of Bias Amplification. *arXiv preprint arXiv:2201.11706* (2022).
- [6] Song Han, Jeff Pool, John Tran, and William Dally. 2015. Learning both weights and connections for efficient neural network. In *Advances in Neural Information Processing Systems (NeurIPS)*.
- [7] Sebastian Hartwig, Dominik Engel, Leon Sick, Hannah Kniesel, Tristan Payer, Poonam Poonam, Michael Glöckler, Alex Bäuerle, and Timo Ropinski. 2025. A Survey on Quality Metrics for Text-to-Image Generation. *arXiv:2403.11821 [cs.CV]* <https://arxiv.org/abs/2403.11821>
- [8] Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. 2022. CLIPScore: A Reference-free Evaluation Metric for Image Captioning. *arXiv:2104.08718 [cs.CV]* <https://arxiv.org/abs/2104.08718>
- [9] Martin Heusel, Hubert Ramsauer, Thomas Unterthiner, Bernhard Nessler, and Sepp Hochreiter. 2018. GANs Trained by a Two Time-Scale Update Rule Converge to a Local Nash Equilibrium. *arXiv:1706.08500 [cs.LG]* <https://arxiv.org/abs/1706.08500>
- [10] Sara Hooker, Aaron Courville, Gregory Clark, Yann Dauphin, and Andrea Frome. 2019. What Do Compressed Deep Neural Networks Forget? *arXiv preprint arXiv:1911.05248* (2019).
- [11] Eugenia B. Iofinova, Renata Bruintjes, Sara Hooker, Mark Ponomarenko, Matej Sulc, and Damian Borth. 2023. Bias in Pruned Vision Models: In-Depth Analysis and Countermeasures. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.
- [12] Ligong Jin, Yang Zhang, Justin Lazarow, Zhuowen Chen, Song Wang, Ji Wang, Yujie Wang, Yujie Wang, and Yujie Wang. 2024. Effortless Efficiency: Low-Cost Pruning of Diffusion Models. *arXiv preprint arXiv:2412.02852* (2024).
- [13] Woosuk Kwon, Sehoon Kim, Michael W. Mahoney, Joseph Hassoun, Kurt Keutzer, and Amir Gholami. 2022. A Fast Post-Training Pruning Framework for Transformers. *arXiv:2204.09656 [cs.CL]* <https://arxiv.org/abs/2204.09656>
- [14] Kimmo Kärkkäinen and Jungseock Joo. 2019. FairFace: Face Attribute Dataset for Balanced Race, Gender, and Age. *arXiv:1908.04913 [cs.CV]* <https://arxiv.org/abs/1908.04913>
- [15] Xiaofeng Lin, Seungbae Kim, and Jungseock Joo. 2022. FairGRAPE: Fairness-aware GRADient Pruning mEthod for Face Attribute Classification. *arXiv:2207.10888 [cs.CV]* <https://arxiv.org/abs/2207.10888>
- [16] Alexandra Sasha Luccioni, Christopher Akiki, Margaret Mitchell, and Yacine Jernite. 2023. Stable Bias: Analyzing Societal Representations in Diffusion Models. *arXiv preprint arXiv:2303.11408* (2023).
- [17] Dmitry Molchanov, Arsenii Ashukha, and Dmitry Vetrov. 2017. Variational Dropout Sparsifies Deep Neural Networks. In *International Conference on Machine Learning (ICML)*.
- [18] Arsha Nagrani, Shan Yang, Anurag Arnab, Aren Jansen, Cordelia Schmid, and Chen Sun. 2021. Attention bottlenecks for multimodal fusion. *Advances in neural information processing systems* 34 (2021), 14200–14213.
- [19] Jinhui Pang, Xinyun Yang, Xiaoyao Qiu, Zixuan Wang, and Taisheng Huang. 2024. MMAF: Masked Multi-modal Attention Fusion to Reduce Bias of Visual Features for Named Entity Recognition. *Data Intelligence* 6, 4 (2024), 1114–1133.
- [20] Aadarsh Ramamoorthy and Mikhail Yurochkin. 2022. Mitigating Bias in Calibration Error Estimation. *arXiv preprint arXiv:2012.08668* (2022).
- [21] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. 2022. High-Resolution Image Synthesis with Latent Diffusion Models. *arXiv:2112.10752 [cs.CV]* <https://arxiv.org/abs/2112.10752>
- [22] Patrick Schramowski, Manuel Brack, Björn Deiseroth, and Kristian Kersting. 2023. Safe Latent Diffusion: Mitigating Inappropriate Degeneration in Diffusion Models. *arXiv preprint arXiv:2211.05105* (2023).
- [23] Preethi Seshadri, Sameer Singh, and Yanai Elazar. 2023. The Bias Amplification Paradox in Text-to-Image Generation. *arXiv:2308.00755 [cs.LG]* <https://arxiv.org/abs/2308.00755>
- [24] Tony Tang, Khai Jun Zhang, Eugene Chuang, Zhixuan Li, Maolin Wu, Kevin Du, Jonathan Wall, Andrew Chang, and James Zou. 2022. Mitigating Gender Bias in Natural Language Processing: Literature Review. *arXiv preprint arXiv:1906.08976* (2022).
- [25] Angelina Wang and Arvind Narayanan. 2021. Directional Bias Amplification. In *International Conference on Machine Learning (ICML)*.
- [26] Chaoqi Wang, Guodong Zhang, and Roger Grosse. 2020. Picking Winning Tickets Before Training by Preserving Gradient Flow. In *International Conference on Learning Representations (ICLR)*.
- [27] Zhou Wang, A.C. Bovik, H.R. Sheikh, and E.P. Simoncelli. 2004. Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing* 13, 4 (2004), 600–612. doi:10.1109/TIP.2003.819861
- [28] Wei Yin, Chi Zhang, Hao Chen, Zhipeng Cai, Gang Yu, Kaixuan Wang, Xiaozhi Chen, and Chunhua Shen. 2023. Structural Pruning for Diffusion Models. *arXiv preprint arXiv:2305.10924* (2023).
- [29] Yang Zhang, Justin Lazarow, Zhuowen Chen, Song Wang, Ji Wang, Yujie Wang, Yujie Wang, Yujie Wang, and Yujie Wang. 2024. EcoDiff: Efficient Pruning for Diffusion Models. *arXiv preprint arXiv:2412.02852* (2024).
- [30] Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints. In *Conference on Empirical Methods in Natural Language Processing (EMNLP)*.