BIASMAP: Can Cross-Attention Uncover Hidden Social Biases?

Rajatsubhra Chakraborty[1] **Xujun Che**[1] **Depeng Xu**[1] **Cori Faklaris**[1]
Xi Niu[1] **Shuhan Yuan**[2]

[1] University of North Carolina at Charlotte [2] Utah State University

{rchakra6,xche,dxu7,cfaklari,xniu2}@charlotte.edu,Shuhan.Yuan@usu.edu

Abstract

Bias discovery is critical for black-box generative models, especially text-to-image (TTI) models. Existing works predominantly focus on output-level demographic distributions, which do not necessarily guarantee concept representations to be disentangled post-mitigation. We propose BiasMap, a model-agnostic framework for uncovering latent concept-level representational biases in stable diffusion models. BiasMap leverages cross-attention attribution maps to reveal structural entanglements between demographics (e.g., gender, race) and semantics (e.g., professions) concepts going deeper into representational bias within the image generation. Using attribution maps of these concepts, we quantify the spatial entanglement via Intersection over Union (IoU), offering a lens into bias that remains hidden in the individual generation process. Our findings show that existing fairness interventions may reduce the output distributional gap but often fail to disentangle concept-level coupling, which is identifiable through our bias discovery method.

1. Introduction

"***With great power, comes great responsibility.***" While the iconic quote from Spider-Man's Uncle Ben Parker was meant for superheroes, it applies just as equivalently to generative models. Stable Diffusion (SD) models [5, 22, 27] hold significant power in creating highly realistic images from input text. However, much like Spider-Man's webs, their outputs are often entangled with inherent biases [37] that frequently go unnoticed. Recent SD models achieve their impressive capabilities by learning statistical patterns from massive internet-sourced datasets comprising billions of images and captions [29]. Yet, these datasets inherently reflect societal biases [2, 6], implicitly perpetuating or amplifying stereotypes involving sensitive demographic attributes such as **gender** and **race** . Such biases pose significant ethical and fairness challenges, as these generative models actively shape public perception. Moreover, due to SD's internal opacity [32], precisely identifying or addressing biases at a representational level remains difficult. Generally, biases originate from two primary sources: data-level imbalances in training sets [15, 31] and latent representational entanglements [42] which are hidden internal correlations between demographic (e.g., **gender** , **race**) and semantic (e.g., **professions**) concepts learned implicitly. Even when data-level issues are corrected, latent entanglements often persist, subtly embedding stereotypes conceptually [14]. We explicitly define *latent entanglement* as internal representational overlaps between demographic and semantic concepts, signifying implicit associations that remain independent of explicit textual conditioning.

Prior works [1, 16] have primarily focused on output-level observation in SD, examining skews in demographic distributions at the generated image level. While these approaches provide valuable insights, they offer limited understanding of the internal representational structures that underpin these biases. Recently, some efforts [10, 17] have been made to inspect biases within the diffusion process itself. However, these studies lack fine-grained spatial-level indicators, offering no direct method to identify precisely which regions or pixels are impacted by bias or how deeply demographic concepts become entangled with semantic attributes. To overcome these limitations and move beyond superficial, output-level bias audits, it is essential to inspect and quantify the latent representational entanglement spatially within generative models. A deeper understanding of how demographic concepts intertwine internally with semantic roles would enable more targeted and effective bias mitigation interventions, going beyond merely adjusting output distributions to structurally addressing biases at the representational level. Therefore, our work explicitly targets this crucial research gap and poses the following central research questions:

RQ1: *How can we leverage attribution mapping to explain the source of bias for generation in SD?*

RQ2: *How do we quantify the bias in SD in the form of concept entanglement and evaluate existing SD models?*

To address these questions, we propose **BiasMap**, a model-agnostic framework utilizing cross-attention attribution maps to quantify latent representational entanglements in text-to-image diffusion models. Our primary contributions include:

- A novel block-wise localization method that precisely identifies and quantifies representational entanglement between demographic attributes and semantic concepts.
- The introduction of Intersection-over-Union (IoU) as a metric for effectively quantifying concept-level biases
- Empirical evidence demonstrating that mIoU provides deeper insights into structural biases, complementing traditional distribution-based metrics to highlighting the necessity of addressing latent representational biases beyond mere demographic parity.

2. BiasMap

2.1. Preliminary

Let f denote a stable diffusion model. Given a prompt P and noise $\mathbf{z}$, the model generates the corresponding image $\mathbf{I} = f(P, \mathbf{z})$ with shape $W \times H \times C$.

In generative TTI models, cross-attention integrates text into image synthesis. Open-Vocabulary Attention Maps (OVAM) [19] attribute spatial influence to arbitrary concepts, even those absent from input prompts. For an arbitrary concept a, which does not need to be in the original prompt P used to generate the image $\mathbf{I}$, OVAM generates an attention attribution map $M_a(\mathbf{I})$ to interpret the spatial region related to the concept a. To construct OVAM, the attribution prompt P' with $a \in P'$ is converted by CLIP encoder as $X' \in \mathbb{R}^{d_E \times d_{X'}}$, where d_E is the embedding dimension and $d_{X'}$ is the number of tokens. Without loss of generality, the concept a is expressed as a single token a. For generating the open-vocabulary attention matrices for even concept $a \notin P$, OVAM uses $\ell_K^{(i)}$ as key projection at each block i to compute the attribution keys: $K_i' = \ell_K^{(i)}(X')$. During denoising, pixel-space queries are extracted at block i, timestep t: $Q_{i,t} = \ell_Q^{(i)}(h_{i,t})$, where $h_{i,t}$ is the i-th convolutional block output at time step t and $\ell_Q^{(i)}$, is learned projection at each block i. The cross-attention matrix $A \in \mathbb{R}^{W^{(i)} \times H^{(i)} \times d_H^{(i)} \times d_{X'}}$ is computed for each block i and time step t:

$$A(Q_{i,t}, K_i') = \text{softmax}\left(\frac{Q_{i,t}K_i'^{\top}}{\sqrt{d}}\right), \tag{1}$$

where d is the query/key dimensionality, $W^{(i)} \times H^{(i)}$ is the reduced latent space shape at block i, $d_H^{(i)}$ is the number of attention heads at block i.

To generate the attribution map $M_a(\mathbf{I})$, OVAM aggregates the matrices across blocks, timestamps, and attention heads for the slices associated with token a:

$$M_a(\mathbf{I}) = \sum_{i,t,l} \text{resize}\left(A_{l,a}(Q_{i,t}, K_i')\right) \in \mathbb{R}^{W \times H}, \tag{2}$$

where $A_{h,k}$ refers to the slice associated with the l-th attention head and token a, and resize$(\cdot)$ normalizes resolution by bilinear interpolation. The map $M_a(\mathbf{I}) \in \mathbb{R}^{W \times H}$ localizes token influence for concept probing. When both P and P' are identical, the heatmaps are equivalent to directly extracting and aggregating the cross-attention matrices computed during image synthesis.

2.2. Methodology

To interpret biased concept association during image synthesis, we propose **BiasMap** , as seen in Figure 1 which spatially localizes concept entanglement during image generation via concept attribution maps.

Bias Localization. To evaluate the generation of $\mathbf{I} = f(P, \mathbf{z})$, we define two attribution prompts with embeddings P_a' and P_b' containing concepts a and b, respectively. In the context of bias discovery, concept a denotes the **demographics** (e.g., gender or **race**) and concept b denotes the **semantics** (e.g., **profession**). In a common TTI setting explored in previous works, $a \notin P$ and $b \in P$. Using Eq. 2, we compute the aggregated attribution maps M_a and H_b indicating spatial attribution for each concept. We model **concept entanglement** as the similarity of cross-attention attribution maps in the pixel space between two concepts. More specifically, we focus on the attribution maps of the high cross-attention regions in the pixel space. We localize the high attention regions with at a threshold τ and generate binary masks:

$$\bar{M}_a[x, y] = \begin{cases} 1, & \text{if } M_a[x, y] \geq \tau, \\ 0, & \text{otherwise,} \end{cases} \tag{3}$$

where $[x, y]$ denotes the 2D spatial coordinates in the attention heatmap of resolution $W \times H$, and $M_a[x, y] \in [0, 1]$. This yields binary masks $\bar{M}_a$ and $\bar{M}_b$ representing regions most influenced by the respective concepts.

We compute the **Intersection over Union (IoU)** between these masks to quantify entanglement:

$$\text{IoU}(\bar{M}_a, \bar{M}_b) = \frac{\sum_{x,y} \bar{M}_a[x, y] \cdot \bar{M}_b[x, y]}{\sum_{x,y} \max(\bar{M}_a[x, y], \bar{M}_b[x, y])}. \tag{4}$$

Intuition

If **concept entanglement** exists between demographics and semantics, the attention maps should have **substantial intersection** over spatial regions. That is, the same pixels are influenced by both concepts during generation.

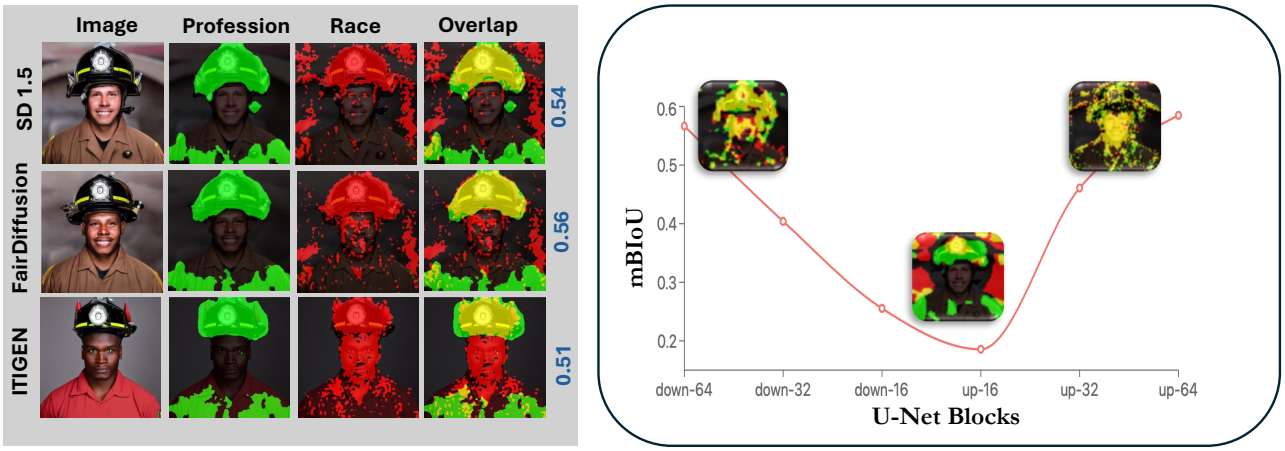

Figure 1. **BiasMap Visualization and Block-wise Entanglement Analysis.** Left panels show BiasMap visualizations across three models (SD 1.5, FairDiffusion, ITIGEN) with attention masks for profession(firefighter), race, and their overlaps (yellow). Right panels display block-level analysis of spatial entanglement (measured by mBIoU) between: profession and race tokens across critical UNet blocks. The visualizations highlight attention overlap at different resolutions, particularly at down-16×16 (middle network blocks), and up-64×64 and down-64×64 blocks (early encoding and final decoding blocks), revealing where demographic-semantic concept entanglement occurs most prominently in the diffusion process. IoU scores are shown in blue.

Lower IoU indicates better separation of demographics and semantics concepts in image generation.

Block-wise Bias Localization. Eq. 2 shows the aggregated attribution map across all blocks. Since concepts are generated in different blocks, we further dive into the block-wise attribution map at block i:

$$M_a^{(i)}(\mathbf{I}) = \sum_{t,l} \text{resize}\left(A_{l,a}(Q_{i,t}, K_i')\right) \in \mathbb{R}^{W\times H}. \quad (5)$$

Similarly, we obtain the block-wise binary masks $\bar{M}_a^{(i)}$ and $\bar{M}_b^{(i)}$ for two concepts and compute the **Block-wise Intersection over Union (BIoU)** to analyze at what depth entanglement occurs.

$$\text{BIoU}^{(i)}(\bar{M}_a^{(i)}, \bar{M}_b^{(i)}) = \frac{\sum_{x,y} \bar{M}_a^{(i)}[x,y] \cdot \bar{M}_b^{(i)}[x,y]}{\sum_{x,y} \max(\bar{M}_a^{(i)}[x,y], \bar{M}_b^{(i)}[x,y])}.$$

For each $\mathbf{I}$, an average BIoU over all blocks is computed.

$$\text{BIoU}(a,b) = \frac{1}{N}\sum_{i=1}^{N} \text{BIoU}^{(i)}(\bar{M}_a^{(i)}, \bar{M}_b^{(i)}), \quad (6)$$

where N is the number of blocks.

Difference from Risk Difference. Previous works only focus on group fairness in generation output distribution. The **Risk Difference (RD)**, defined as $\text{RD}(a_1, a_2) = |\Pr(a_1) - \Pr(a_2)|$, where a_1, a_2 are different demographic groups (e.g., "male" and "female"), measures group-level bias through demographic parity across samples, addressing distributional fairness. Conversely, IoU or BIoU quantifies representational entanglement by measuring spatial co-activation patterns for each individual generation. it reveals how demographics become structurally coupled with semantics concept in the latent space. We can also extend concept entanglement to a group setting. We evaluate the mIoU (or mBIoU) for different images generated by f with the same instruction prompt P but different noises $\mathbf{z}$.

3. Evaluation Setup

Setup. We evaluate bias discovery on 20 occupation prompts selected from the US Bureau of Labor Statistics [36]. Generation prompts P follow the template: "A photo of the face of a [profession]". We evaluate bias against **race** and gender separately. The attribution prompts P' for BiasMap analysis use: "A photo of the face of a [profession] and [race/gender]". For each **profession** , we generate 100 images using Stable Diffusion v1.5 (SD 1.5) and two debiasing models based on RD: FairDiffusion (FG) [7] and ITI-GEN (IG) [41]. For race debiasing, FD modifies the text prompt to alternate the specified race for roughly half the samples, ensuring a balance between white and black depictions. IG operates via latent guidance, conditioning generation on skin tone using the six-point Fitzpatrick scale (1=lightest, 6=darkest) to produce a spectrum of ethnic appearances. For gender debiasing, FD alternates prompts between male and female subjects, whereas IG applies latent guidance to balance masculine vs. feminine features without explicit prompt keywords. We quantify group fairness using RD, where for race, we compute $|\Pr(\text{"white"}) - \Pr(\text{"black"})|$. For gender, we calculate $|\Pr(\text{"male"}) - \Pr(\text{"female"})|$. We use a CLIP ViT-L/14

Table 1. Quantitative results on gender bias. Lower mIoU indicates better disentanglement.

Method	RD ↓	mIoU ↓	mBIoU ↓
SD1.5	0.59	0.3634	**0.4291**
FairDiffusion (FD)	**0.14**	0.3932	0.4591
ITIGEN (IG)	0.17	**0.3579**	0.4391

Table 2. Quantitative results on **race** bias. Lower mIoU indicates better disentanglement.

Method	RD ↓	mIoU ↓	mBIoU ↓
SD1.5	0.75	**0.4135**	**0.4391**
FairDiffusion	**0.21**	0.4384	0.4394
ITIGEN	0.22	0.4034	0.4498

model for demographics classification [23], with race classifications on white, black, Hispanic, and Asian. We quantify concept entanglement using mIoU and mBIoU over 100 images. We set the threshold τ in Eq. 3 to 70% to generate our binarized heatmaps. All experiments were run on a single NVIDIA A100 GPU.

4. Quantitative Evaluation

Performance Analysis. Tables 1 and 2 reveal two critical observations: (1) Both FairDiffusion and ITIGEN effectively mitigate demographic disparities, significantly reducing RD by 70-75% for both gender and race attributes across occupational contexts based on t-test, (2) Despite this distributional improvement, the mIoU remains virtually unchanged or even slightly increases post intervention (no significant changes). This is consistent across both demographic concepts and both variants of mIoU. Extended experiment results are included in Supp. C to F.

Structural Implications. The observed divergence between demographic parity and persistent concept entanglement indicates an interesting limitation in present debiasing methods. While interventions successfully rebalance output distributions, they fail to disentangle the underlying representational structures where demographic and semantic concepts become spatially coupled in the model's internal representations. This observation suggests that effective bias mitigation requires addressing not only the class of the output but also how concepts are structurally associated within the generation process itself.

5. Findings

5.1. RQ1: Latent Bias Beyond Outputs

Our results reveal that SD's internal representations encode demographic–semantic entanglements beyond what output-level audits capture. Even when prompts do not specify gender or **race** , cross-attention maps show substantial spatial overlap (mIoU) between demographic tokens and certain **professions** . For instance, *nurse* strongly co-activates in heatmap with gender in the early attention blocks (refer to Supp. B), aligning with stereotypical output biases.We notice this bias emerges within the U-Net, well before final image generation or any demographic balancing. Hence, reducing output skew alone, say by adjusting sampling, does not necessarily alter how the model internally associates demographic attributes with semantic ones.

Key Finding 1

The inception of bias is in the U-Net, way before final image generation. Activated heatmaps highlight the source of bias lies in early high-resolution blocks and resurfaces in final decoding blocks.In particular we notice high mBIoU in the first downsampling 64×64 and last upsampling 64×64 blocks, showing a convex non-monotonic trend akin to the U-shape of the subnetwork.

5.2. RQ2: Concept Entanglement Quantified

In our study we reduce the abstraction in the idea of concept entanglement to a quantifiable paradigm using cross-attention based activation maps as discussed in Section 2. In particular we represent entanglement as pixel-wise overlap in attention outputs for our two concepts, using mIoU as the representing metric. For our distributional metric we propose Risk Difference (RD) to denote group-level bias. This dual-measurement paradigm enables both macro-level fairness evaluation and micro-level interpretability of embedded demographic markers, providing a more comprehensive framework for generative model bias analysis.

Key Finding 2

Output-level parity does not imply latent fairness. SD's cross-attention confirms that **professions** remain gendered or racialized within the network, indicating that mere output balancing fails to remove deeper conceptual bias.

6. Conclusion

BiasMap uncovers latent conceptual biases in Stable Diffusion by measuring spatial entanglement between demographics and semantic concepts using cross-attention. We show that distribution based mitigation is not sufficient by highlighting persisting entangled concept correlations even in the debiased models. Currently, we focus on single-axis bias and do not explore intersectional biases where multiple demographic attributes co-occur. This leaves open questions about how compounding identities influence internal representations. We aim to develop mitigation techniques that directly reduce entanglement during generation by guiding models toward concept disentanglement.

Acknowledgements

This work was supported in part by National Science Foundation 2348391.

References

[1] Nouar AlDahoul, Talal Rahwan, and Yasir Zaki. Ai-generated faces influence gender stereotypes and racial homogenization. *arXiv preprint arXiv:2402.01002*, 2024. 1

[2] Abeba Birhane, Sepehr Dehdashtian, Vinay Prabhu, and Vishnu Boddeti. The dark side of dataset scaling: Evaluating racial classification in multimodal models. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency*, pages 1229–1244, 2024. 1

[3] Moreno D'Incà, Elia Peruzzo, Massimiliano Mancini, Dejia Xu, Vidit Goel, Xingqian Xu, Zhangyang Wang, Humphrey Shi, and Nicu Sebe. Openbias: Open-set bias detection in text-to-image generative models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12225–12235, 2024. 2

[4] Alexei A Efros and Thomas K Leung. Texture synthesis by non-parametric sampling. In *Proceedings of the seventh IEEE international conference on computer vision*, pages 1033–1038. IEEE, 1999. 1

[5] Patrick Esser, Sumith Kulal, Andreas Blattmann, Rahim Entezari, Jonas Müller, Harry Saini, Yam Levi, Dominik Lorenz, Axel Sauer, Frederic Boesel, et al. Scaling rectified flow transformers for high-resolution image synthesis. *arXiv preprint arXiv:2403.03206*, 2024. 1

[6] Simone Fabbrizzi, Symeon Papadopoulos, Eirini Ntoutsi, and Ioannis Kompatsiaris. A survey on bias in visual datasets. *Computer Vision and Image Understanding*, 223: 103552, 2022. 1

[7] Felix Friedrich, Manuel Brack, Lukas Struppek, Dominik Hintersdorf, Patrick Schramowski, Sasha Luccioni, and Kristian Kersting. Fair diffusion: Instructing text-to-image generation models on fairness, 2023. 3, 2

[8] 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. 1

[9] David J Heeger and James R Bergen. Pyramid-based texture analysis/synthesis. In *Proceedings of the 22nd annual conference on Computer graphics and interactive techniques*, pages 229–238, 1995. 1

[10] Eunji Kim, Siwon Kim, Rahim Entezari, and Sungroh Yoon. Unlocking intrinsic fairness in stable diffusion. *arXiv preprint arXiv:2408.12692*, 2024. 1

[11] Diederik P Kingma, Max Welling, et al. Auto-encoding variational bayes, 2013. 1

[12] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015. 1

[13] Hang Li, Chengzhi Shen, Philip Torr, Volker Tresp, and Jindong Gu. Self-discovering interpretable diffusion latent directions for responsible text-to-image generation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 12006–12016, 2024. 1

[14] Francesco Locatello, Gabriele Abbati, Thomas Rainforth, Stefan Bauer, Bernhard Schölkopf, and Olivier Bachem. On the fairness of disentangled representations. *Advances in neural information processing systems*, 32, 2019. 1

[15] Alexandra Sasha Luccioni, Christopher Akiki, Margaret Mitchell, and Yacine Jernite. Stable bias: Analyzing societal representations in diffusion models, 2023. 1

[16] Sasha Luccioni, Christopher Akiki, Margaret Mitchell, and Yacine Jernite. Stable bias: Evaluating societal representations in diffusion models. *Advances in Neural Information Processing Systems*, 36:56338–56351, 2023. 1

[17] Abhishek Mandal, Susan Leavy, and Suzanne Little. Generated bias: Auditing internal bias dynamics of text-to-image generative models. *arXiv preprint arXiv:2410.07884*, 2024. 1

[18] Elman Mansimov, Emilio Parisotto, Jimmy Lei Ba, and Ruslan Salakhutdinov. Generating images from captions with attention. *arXiv preprint arXiv:1511.02793*, 2015. 1

[19] Pablo Marcos-Manchón, Roberto Alcover-Couso, Juan C SanMiguel, and Jose M Martínez. Open-vocabulary attention maps with token optimization for semantic segmentation in diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9242–9252, 2024. 2, 1

[20] Hadas Orgad, Bahjat Kawar, and Yonatan Belinkov. Editing implicit assumptions in text-to-image diffusion models. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7053–7061, 2023. 2

[21] Rishubh Parihar, Abhijnya Bhat, Abhipsa Basu, Saswat Mallick, Jogendra Nath Kundu, and R Venkatesh Babu. Balancing act: distribution-guided debiasing in diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 6668–6678, 2024. 2

[22] Dustin Podell, Zion English, Kyle Lacey, Andreas Blattmann, Tim Dockhorn, Jonas Müller, Joe Penna, and Robin Rombach. Sdxl: Improving latent diffusion models for high-resolution image synthesis. In *The Twelfth International Conference on Learning Representations*, 2025. 1

[23] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. Learning transferable visual models from natural language supervision, 2021. 4

[24] Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. Zero-shot text-to-image generation. In *International conference on machine learning*, pages 8821–8831. Pmlr, 2021. 1

[25] Aditya Ramesh, Prafulla Dhariwal, Alex Nichol, Casey Chu, and Mark Chen. Hierarchical text-conditional image generation with clip latents. *arXiv preprint arXiv:2204.06125*, 1 (2):3, 2022. 1

[26] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models, 2021. 1

[27] 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*, pages 10684–10695, 2022. 1

[28] Chitwan Saharia, William Chan, Saurabh Saxena, Lala Li, Jay Whang, Emily L Denton, Kamyar Ghasemipour, Raphael Gontijo Lopes, Burcu Karagol Ayan, Tim Salimans, et al. Photorealistic text-to-image diffusion models with deep language understanding. *Advances in neural information processing systems*, 35:36479–36494, 2022. 1

[29] Christoph Schuhmann, Romain Beaumont, Richard Vencu, Cade Gordon, Ross Wightman, Mehdi Cherti, Theo Coombes, Aarush Katta, Clayton Mullis, Mitchell Wortsman, Patrick Schramowski, Srivatsa Kundurthy, Katherine Crowson, Ludwig Schmidt, Robert Kaczmarczyk, and Jenia Jitsev. Laion-5b: An open large-scale dataset for training next generation image-text models, 2022. 1

[30] Preethi Seshadri, Sameer Singh, and Yanai Elazar. The bias amplification paradox in text-to-image generation. *arXiv preprint arXiv:2308.00755*, 2023. 1

[31] Preethi Seshadri, Sameer Singh, and Yanai Elazar. The bias amplification paradox in text-to-image generation, 2023. 1

[32] Yingdong Shi, Changming Li, Yifan Wang, Yongxiang Zhao, Anqi Pang, Sibei Yang, Jingyi Yu, and Kan Ren. Dissecting and mitigating diffusion bias via mechanistic interpretability. *arXiv preprint arXiv:2503.20483*, 2025. 1, 2

[33] Jascha Sohl-Dickstein, Eric Weiss, Niru Maheswaranathan, and Surya Ganguli. Deep unsupervised learning using nonequilibrium thermodynamics. In *International conference on machine learning*, pages 2256–2265. pmlr, 2015. 1

[34] Raphael Tang, Linqing Liu, Akshat Pandey, Zhiying Jiang, Gefei Yang, Karun Kumar, Pontus Stenetorp, Jimmy Lin, and Ferhan Ture. What the daam: Interpreting stable diffusion using cross attention. In *The 61st Annual Meeting Of The Association For Computational Linguistics*, 2023. 1

[35] Michael Toker, Hadas Orgad, Mor Ventura, Dana Arad, and Yonatan Belinkov. Diffusion lens: Interpreting text encoders in text-to-image pipelines. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9713–9728, 2024. 1

[36] U.S. Bureau of Labor Statistics. Employed persons by detailed occupation, sex, race, and hispanic or latino ethnicity, 2025. Accessed: 2025-04-04. 3

[37] Adriana Fernández de Caleya Vázquez and Eduardo C Garrido-Merchán. A taxonomy of the biases of the images created by generative artificial intelligence. *arXiv preprint arXiv:2407.01556*, 2024. 1

[38] Jialu Wang, Xinyue Gabby Liu, Zonglin Di, Yang Liu, and Xin Eric Wang. T2iat: Measuring valence and stereotypical biases in text-to-image generation. *arXiv preprint arXiv:2306.00905*, 2023. 1

[39] Yankun Wu, Yuta Nakashima, and Noa Garcia. Revealing gender bias from prompt to image in stable diffusion. *Journal of Imaging*, 11(2):35, 2025. 2

[40] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. *arXiv preprint arXiv:2206.10789*, 2(3):5, 2022. 1

[41] Cheng Zhang, Xuanbai Chen, Siqi Chai, Henry Chen Wu, Dmitry Lagun, Thabo Beeler, and Fernando De la Torre. ITI-GEN: Inclusive text-to-image generation. In *ICCV*, 2023. 3, 2

[42] Mengdan Zhu, Raasikh Kanjiani, Jiahui Lu, Andrew Choi, Qirui Ye, and Liang Zhao. Latentexplainer: Explaining latent representations in deep generative models with multimodal large language models, 2025. 1

BIASMAP: Can Cross-Attention Uncover Hidden Social Biases?

Supplementary Material

The supplementary material is divided as follows: Section A discusses existing literature in depth regarding image generation, bias discovery, interpretability and bias mitigation. Section B provides a deeper discussion into blockwise entanglement analysis in the denoising step of diffusion. Section C highlights extended experiments and their quantitative results. In our final sections, we strengthen our proposal of mIoU as a metric: Section E validates mIoU as a correct measure for understanding entanglement, and Section F discusses the faithfulness of mIoU.

A. Related Works

Image Synthesis and Stable Diffusion. Image synthesis at its earliest stages relied heavily on deterministic algorithms and feature engineering [4, 9] limiting realism and flexibility. The introduction of deep learning [12] brought about new paradigms like such as Variational Autoencoders (VAEs) [11] and Generative Adversarial Networks (GANs) [8]. While these significantly improved generation quality, [11] used objective functions that often led to blurry images, and [8] faced challenges such as training instability and mode collapse. The earliest works of text-to-image (TTI) used VAEs with text sequences [18]. At this point, the idea of diffusion came into being [33], leading to the approach of generation being treated as a process of denoising , starting from pure noise. Latent Diffusion Models introduced an efficient mechanism by operating in a compressed latent space, significantly reducing computational requirements without compromising on image fidelity. Stable Diffusion [26] emerged as the foundational model for TTI synthesis. The earliest versions of SD were trained on 512 X 512 images with a CLIP ViT-L/14 text encoder and later, with a CLIP ViT-H/14 variant. SDXL [22] featured a base model with 3.5B parameters, significantly larger than its predecessors and introduced native support for 1024 X1024 images with improved generation of complex features. [5] stands as the latest SD series of models offering variants upto 8B parameters.

Other TTI models. OpenAI's DALL-E [24] introduced an approach using the transformer architecture with a discrete variational autoencoder (dVAE). Building upon its predecessor, DALL-E 2 [25] introduced a two stage framework having a prior network generating CLIP image embeddings from text, followed by a decoder that produces images conditioned on these embeddings. Incorporating CLIP notably improved the model's understanding of semantic concepts resulting to better generation quality. In terms of frontier models, Google introduced Imagen [28] utilized a pretrained text encoder to process textual descriptions, which were then used to condition a series of diffusion models leading to image generation in a cascaded manner. This led to production of photorealistic images and nuanced language understanding. Google also introduced the Pathways Autoregressive Text-to-Image model (Parti) [40] , approaching TTI generation as a sequence-to-sequence problem akin to machine translation. It employed an autoregressive transformer model that generates sequences of image tokens based on input text, facilitating the creation of complex and content-rich images. This method enabled Parti to handle intricate compositions and incorporate extensive world knowledge into the generated imagery.

Interpretability and Bias Discovery in SD. The interpretability of diffusion models, particularly those in the SD model family, has been the focus of recent studies aiming to elucidate the internal mechanisms of generation. This also aims to discover internal biases withiin the TTI models. Diffusion Attentive Attribution Maps (DAAM) [34] generates pixel-level attribution maps by upscaling and aggregating cross-attention scores from SD's denoising network. Diffusion Lens[35] focuses on the text encoder component of text-to-image pipelines. By generating images conditioned on intermediate text representations, it provides insights into how textual information is processed and utilized during image synthesis. This method sheds light on the compositional understanding of complex prompts. However, it primarily examines the text encoder in isolation, potentially overlooking the interplay between text and image components in the diffusion process. Open Vocabulary Attention Maps (OVAM) [19]is a training-free method that enables the generation of attention maps for arbitrary words, extending beyond the original text prompts used in image synthesis. It includes a lightweight token optimization process that enhances the accuracy of these attention maps with minimal supervision.

Recent bias discovery methods [13, 30, 32, 38] for SD models primarily focus on intermediate representational observations to gain insights into inherent biases within the model. [13] proposed a self-supervised technique to extract interpretable latent directions corresponding to semantic attributes without labeled supervision, enabling attribute disentanglement and surfaced latent bias axes. However, its focus was restricted to binary-aligned semantic traits and not arbitrary concepts. [30] introduces a bias amplification paradox framework, comparing the distribution of attributes in generated image against those implied in train-

ing captions. This revealed that SD disproportionately amplified biases even when prompts are neutral, underscoring the role of both training data priors and model prompt alignment mismatches. Recently, OpenBias [3] introduced a flexible pipeline for open-set bias discovery without requiring pre-specified demographic categories. Leveraging a combination of image generation, large vision-language models (VLMs), and question-answering modules, OpenBias identified both known and emergent biases across diverse prompts. [39] highlighted that gendered associations not only influence face and body generation but also bias object placement and compositional structure, suggesting entrenched priors in both the text encoder and image generator. To our knowledge, the most recent work is [32] which uncovered localized structures in the generative process responsible for encoding bias-correlated concepts and proposed patching interventions to mitigate these pathways, enabling bias-aware control without architectural retraining.

Bias Mitigation. Fair Diffusion [7] introduced a strategy that allows users to guide model outputs via human instructions, effectively adjusting biases to achieve desired demographic representations by leveraging concepts captured during training . Inclusive Text-to-Image Generation (ITI-GEN) [41] proposed using reference images to guide the generation process, ensuring the inclusion of diverse attributes without necessitating model fine-tuning . Concurrently, Text-to-Image Model Editing (TIME) [20] was developed to modify implicit assumptions in diffusion models by updating cross-attention layers based on source and destination prompts, allowing for the correction of outdated or biased assumptions .Recently, [21] introduced distribution guidance to condition the reverse diffusion process on a reference attribute distribution, effectively reducing biases without additional data or model retraining . These methodologies represent significant advancements in promoting fairness and inclusivity in generated content by offering diverse strategies to mitigate biases within text-to-image diffusion models.

B. Block-level Analysis

We present a detailed quantitative block-level analysis of conceptual entanglement across the U-Net architecture in diffusion models. We focus specifically on examining five representative professions as a case study: **nurse, firefighter, journalist, chef, and doctor**, analyzing how demographic-semantic concept coupling manifests at different resolutions throughout the network. Figures 2 and 3 present mean Block-wise Intersection-over-Union (mBIoU) measurements across critical U-Net blocks. As highlighted in Key Finding 2 in Section 5.1, the observed pattern

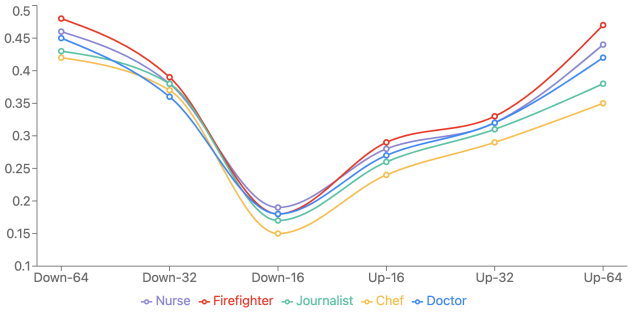

Figure 2. Profession -Gender Concept Entanglement (mBIoU) across UNet Blocks.

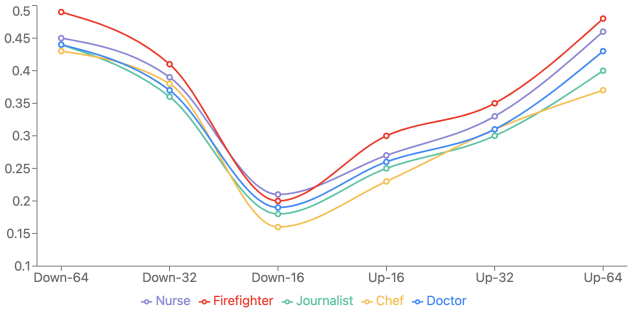

Figure 3. Profession -Race Concept Entanglement (mBIoU) across UNet Blocks.

follows a characteristic U-shaped distribution across network depth, with significantly higher entanglement at high-resolution blocks and lower entanglement at intermediate representations. This suggests that demographic-semantic associations are encoded primarily during initial feature extraction and final image synthesis stages. In further subsections, we provide detailed observations into U-Net layers over profession -gender entanglement with mBIoU values reported as shown in Figure 2.

B.1. High-Resolution Encoding Blocks

The **down-64×64** block exhibits pronounced concept entanglement across all analyzed professions .For professions with strong societal gender associations, such as *nurse*, entanglement reaches 0.46, while *firefighter* shows even higher coupling at 0.48.This indicates that initial feature extraction stages immediately encode demographic attributes as intrinsically linked to profession -based semantics.Notably, our subset of strongly stereotyped professions demonstrate the highest entanglement at this early stage. The *firefighter* profession shows maximal demographic-semantic coupling (0.48), significantly higher than less stereotypically gendered professions like *chef* (0.42).

B.2. Intermediate Representation Blocks

As information flows deeper into the network, we observe progressive disentanglement of demographic and semantic concepts. The **down-32×32** block shows moderate reductions in mBIoU across all **professions** , with values ranging from 0.36 (doctor) to 0.39 (firefighter). Most significantly, the **down-16×16** block corresponding to the network's bottleneck demonstrates substantially reduced entanglement, with mBIoU values dropping to 0.15–0.19 range. This represents a reduction of approximately 60% compared to the initial encoding blocks, suggesting that abstract latent representations partially disentangle demographic attributes from **profession** semantics.

B.3. Generative Upsampling Blocks

In the upsampling phase, we notice that entanglement progressively increases through successive upsampling blocks. Beginning with the **up-16×16** block, mIoU values rise to the 0.24–0.29 range, already showing re-entanglement compared to the bottleneck. The **up-32×32** block continues this trend with further increased coupling (0.29–0.33), while the **up-64×64** block exhibits substantially higher entanglement, particularly for stereotyped **professions** . The *firefighter* **profession** shows the highest terminal entanglement (0.47), closely followed by *nurse* (0.44). This progressive re-entanglement during upsampling suggests that the diffusion model reconstructs demographic-semantic associations during image synthesis, even when these associations were partially disentangled in abstract latent representations.

B.4. Professional Variation in Entanglement Dynamics

Different **professions** exhibit characteristic entanglement signatures across the network architecture. The *firefighter* **profession** consistently shows the highest entanglement at both extremes of the network (0.48 at down-64×64 and 0.47 at up-64×64), suggesting deeply encoded **gender** and **race** associations. The *nurse* **profession** demonstrates the second-highest overall entanglement, with particularly strong coupling during final image synthesis (0.44 at up-64×64) shown in Figure 3. Interestingly, *chef* and *journalist* show more moderate terminal entanglement (0.35 and 0.38 respectively), suggesting potentially weaker but still significant stereotypical associations.

C. Quantitative Results

We show further quantitative results in Table 3 and 4, and highlight some observations apart from the main findings discussed in Section 5. Importantly the trends observed in this section reinforce our original argument presented initially regarding output level distribution not being enough for bias mitigation and/or discovery. We also perform

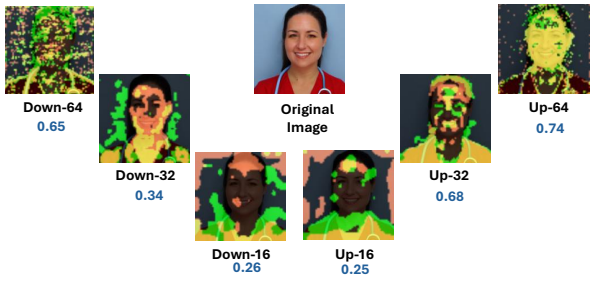

Figure 4. **Profession** -Gender Concept Entanglement (BIoU) across UNet blocks. Note that the yellow mask(s) denote overlap between demographic(**gender**) and semantic(**profession** : Nurse) concepts. BIoU scores are shown in blue.

cross-model validation with the v2 variant of SD and conclude that the trends in bias remain oddly consistent after a marked improvement in architecture design from v1.5. Both discussions highlight the emerging need for better mitigation paradigms to diminish concept-level entanglement within the model.

Asymmetry Between Gender and Race Bias. We observe distinct patterns of conceptual entanglement across demographic dimensions. For instance, *firefighter* exhibits high **gender** entanglement (mIoU = 0.566) but moderate **race** entanglement (mIoU = 0.353), while *musician* shows comparable values for both **gender** (mIoU = 0.519) and **race** (mIoU = 0.466). This suggests that bias dimensions operate independently rather than uniformly across demographic categories.

Counterintuitive Intervention Effects. For certain **professions** , FD paradoxically increases **race** entanglement despite reducing distributional bias. This is most pronounced for *journalist* (mIoU increasing from 0.434 to 0.483, an 11.3% increase) and *artist* (from 0.349 to 0.430, a 23.2% increase). This suggests that intervention methods may inadvertently strengthen internal representational coupling while achieving surface-level distributional fairness.

Disconnect Between Distributional and Representational Bias. Professions demonstrating near-perfect distributional balance often maintain substantial concept entanglement. *Journalist* exhibits near-zero **race** distributional bias (RD = 0.020) yet maintains high race-profession entanglement (mIoU = 0.434). This discrepancy is also evident for *athlete* (gender RD = 0.140, gender mIoU = 0.526), revealing a hidden layer of bias that output-level metrics fail to detect.

Demographic-Specific Entanglement Patterns. The discrepancy between aggregation methods varies markedly

Table 3. Gender based Metrics Comparison. **FD** denotes FairDiffusion and **IG** denotes ITI-GEN.

profession	mIoU ↓			mBIoU ↓			RD ↓		
	SD v1.5	FD	IG	SD 1.5	FD	IG	SD 1.5	FD	IG
architect	**0.490**	0.492	0.492	0.409	**0.408**	0.437	0.940	0.120	**0.100**
artist	**0.493**	0.501	0.501	**0.459**	0.460	0.465	0.860	**0.140**	0.340
athlete	0.526	0.524	0.524	0.481	0.481	0.475	0.140	**0.120**	0.260
cashier	0.443	0.435	0.435	0.389	0.390	**0.386**	0.920	0.180	**0.120**
chef	0.421	0.417	0.417	0.457	0.458	0.462	0.900	0.300	**0.080**
doctor	0.341	**0.336**	**0.336**	0.451	0.451	0.451	0.920	**0.320**	0.360
driver	**0.403**	**0.403**	**0.403**	0.385	**0.384**	0.397	0.900	0.200	**0.020**
engineer	0.440	**0.434**	**0.434**	0.426	0.426	0.454	0.900	0.220	**0.160**
firefighter	0.566	0.572	0.572	0.410	**0.409**	0.417	0.980	0.260	**0.000**
journalist	0.524	0.524	0.524	0.447	0.447	0.457	0.980	0.240	**0.040**
lawyer	0.474	0.475	0.475	0.433	0.433	0.444	0.900	0.240	**0.220**
mechanic	0.430	**0.427**	**0.427**	**0.361**	**0.361**	0.363	0.960	**0.120**	0.180
musician	0.519	0.527	0.527	**0.462**	**0.462**	0.476	0.480	**0.120**	0.460
nurse	**0.338**	**0.338**	**0.338**	0.474	0.474	0.477	0.600	**0.020**	0.580
officer	0.407	**0.394**	**0.394**	0.472	0.471	0.467	0.760	**0.080**	0.140
pilot	0.430	0.406	0.406	0.439	**0.438**	0.442	0.980	0.400	**0.180**
scientist	0.399	**0.393**	**0.393**	**0.458**	**0.458**	0.479	0.980	0.480	**0.400**
teacher	0.414	**0.405**	**0.405**	0.485	0.485	0.508	0.900	**0.100**	0.300
waiter	0.469	0.467	0.467	0.464	0.463	0.470	1.000	0.440	**0.020**
worker	**0.298**	0.299	0.299	**0.461**	0.462	0.482	0.480	**0.080**	0.520

Table 4. Race based Metrics Comparison.**FD** denotes FairDiffusion and **IG** denotes ITI-GEN.

profession	mIoU ↓			mBIoU ↓			RD ↓		
	SD 1.5	FD	IG	SD 1.5	FD	IG	SD 1.5	FD	IG
architect	**0.431**	0.472	0.454	**0.424**	0.426	0.436	0.700	0.160	**0.120**
artist	**0.349**	0.430	0.368	**0.437**	0.438	0.443	0.280	**0.060**	0.100
athlete	0.487	0.505	**0.451**	**0.479**	**0.479**	0.491	0.320	**0.020**	0.240
cashier	0.292	0.315	**0.251**	**0.419**	0.420	0.423	0.540	**0.100**	0.300
chef	0.433	0.455	**0.346**	**0.435**	0.437	0.452	0.780	0.200	**0.180**
doctor	0.362	0.378	**0.327**	0.435	0.434	**0.427**	0.220	**0.140**	0.260
driver	0.214	0.238	**0.164**	**0.449**	0.451	0.471	0.680	**0.020**	0.060
engineer	**0.396**	0.430	0.469	0.417	0.418	**0.414**	0.800	0.200	**0.100**
firefighter	0.353	0.381	**0.328**	0.482	0.482	**0.480**	0.980	**0.040**	0.260
journalist	0.434	0.483	**0.432**	**0.427**	0.428	0.429	**0.020**	0.100	0.120
lawyer	0.370	0.406	**0.366**	0.442	0.443	**0.437**	0.420	0.240	**0.120**
mechanic	0.175	0.173	**0.171**	0.175	0.173	**0.171**	0.880	0.100	**0.020**
musician	**0.466**	0.503	0.475	**0.420**	**0.420**	0.435	0.240	0.120	**0.080**
nurse	0.404	0.419	**0.353**	**0.454**	0.455	0.458	1.000	**0.620**	0.840
officer	**0.347**	0.355	0.351	0.485	**0.484**	0.487	0.880	0.080	**0.040**
pilot	0.335	0.337	**0.331**	0.478	**0.476**	0.488	0.560	**0.020**	0.160
scientist	**0.375**	0.423	0.394	0.433	0.433	**0.432**	0.560	0.220	**0.120**
teacher	**0.334**	0.363	0.383	0.474	**0.473**	0.486	0.440	0.160	**0.140**
waiter	0.370	0.400	**0.319**	0.463	**0.462**	0.471	1.000	0.120	**0.000**
worker	**0.346**	0.398	0.426	**0.346**	0.398	0.426	0.520	**0.100**	**0.100**

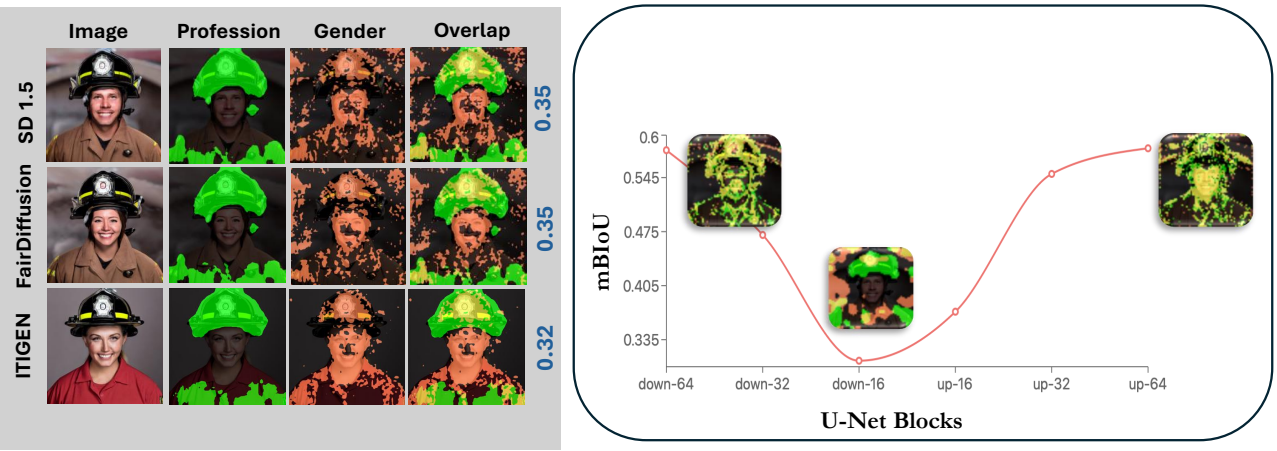

Figure 5. Left panels show BiasMap visualizations for three models (SD 1.5, FairDiffusion, ITIGEN) with attention masks for profession(firefighter), gender, and their overlaps (yellow). Right panels show spatial entanglement (mBIoU) between profession and gender tokens across key UNet blocks, highlighting attention overlap at different resolutions. IoU scores are shown in blue.

across demographic dimensions. For gender bias, differences appear most pronounced in stereotypically gendered **professions** like *firefighter* (gap of 0.156 between layer-wise and block-aggregated values) and *nurse* (gap of 0.136). Conversely, for **race** bias, *chef* shows a minimal gap of 0.002 between measurements, suggesting more uniform distribution of **race** bias across network components.

Resilience of Deeply Entrenched Biases. Professions with extreme initial distributional bias show varying degrees of improvement under interventions, suggesting a ceiling effect in current bias mitigation approaches. For *waiter* with gender bias, we observe perfect distributional bias in SD v1.5 (RD = 1.000), which improves substantially with FD (RD = 0.440) and dramatically with IG (RD = 0.020). However, for *nurse* with **race** bias, we see complete bias in SD v1.5 (RD = 1.000) that only moderately improves with FD (RD = 0.620) and actually worsens with IG (RD = 0.840). Even when distributional metrics improve dramatically (as with *firefighter* gender bias going from RD = 0.980 to RD = 0.000), the entanglement metrics remain high (mIoU = 0.572), indicating that internal representations maintain stereotypical associations despite balanced output.

Summary. The findings validate our hypothesis discussed in Section 1 that simply balancing output distribution might not be enough for mitigation techniques to ensure that bias is no more present. Rather, internal learning representations may increase entanglement of biases as discussed above.

Cross-Model Validation with SDv2. Tables 5 and 6 present compelling evidence for the model-agnostic nature of our BiasMap framework, demonstrating that concept entanglement is a persistent phenomenon across different model architectures and training paradigms. The data reveals that SDv2, despite architectural improvements over SD 1.5, exhibits remarkably similar patterns of bias in both distributional metrics and representational entanglement, validating our approach as a generalizable bias discovery method. Examining SDv2's baseline gender bias in Table 5, we observe that distributional bias (RD) patterns closely mirror those in SD 1.5, with **professions** like *waiter* (RD = 1.000), *firefighter* (RD = 0.980), and *scientist* (RD = 0.980) showing extreme gender imbalance. This consistency across model generations suggests that demographic skews originate primarily from training data biases rather than model-specific architectural features. However, when comparing representational entanglement measurements, SDv2 demonstrates notable differences, with **professions** like *artist* showing substantially lower gender entanglement (mIoU = 0.245) compared to SD 1.5 (mIoU = 0.493). This architectural variation in concept coupling, despite similar distributional biases, confirms that our mIoU measurement captures model-internal representational properties that are distinct from output-level statistics.

The effectiveness of FairDiffusion intervention remains consistent across model generations, with comparable reductions in distributional bias despite architectural differences. For instance, firefighter gender bias decreases from 0.980 to 0.260 in both SD 1.5 and SDv2, demonstrating that text-prompt interventions operate effectively regardless of model architecture. However, the intervention's impact on

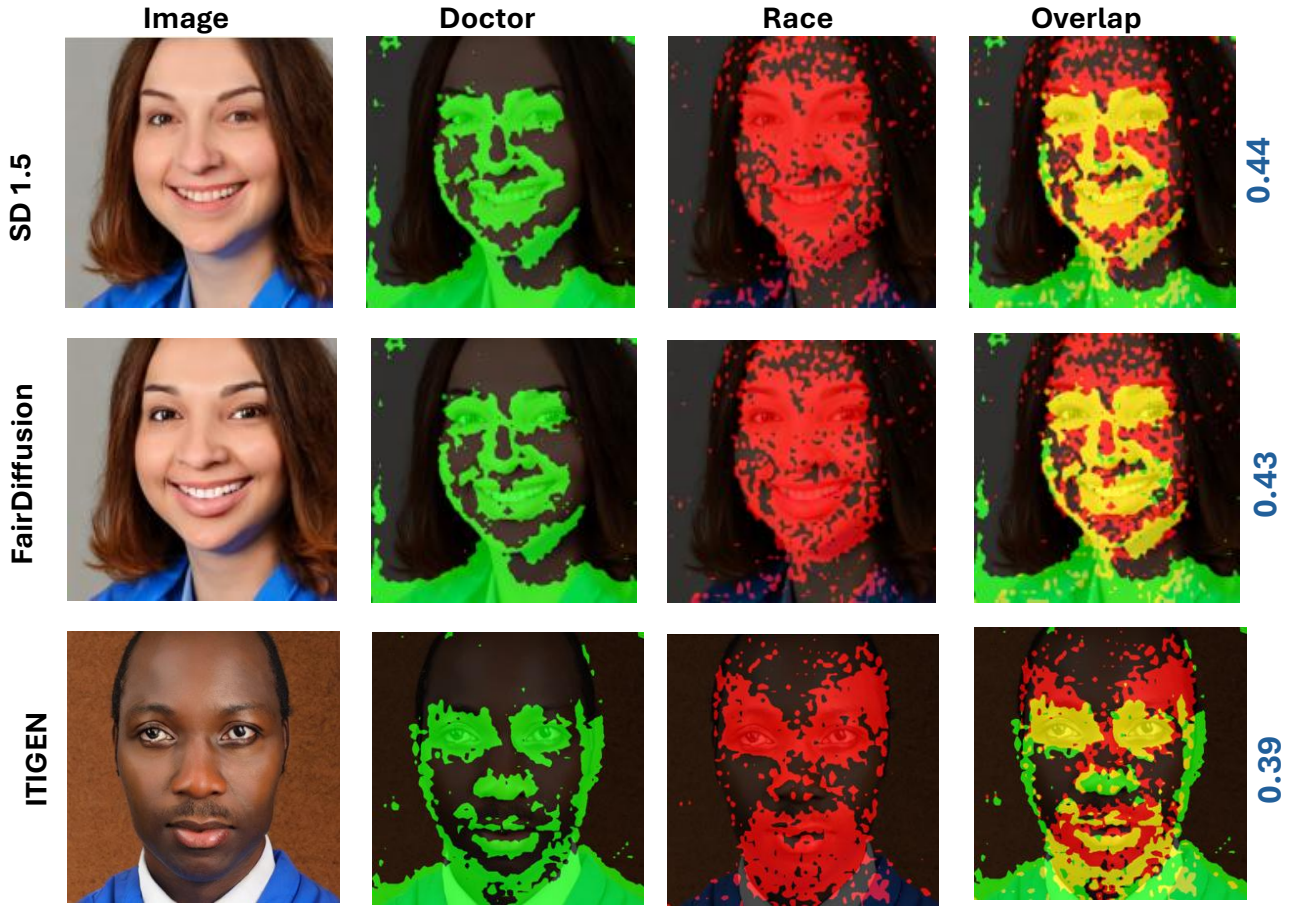

Figure 6. An overview of spatial concept overlap using a generated image of a doctor. Heatmaps highlighted show the effects of attributes: Doctor, Race and Overlapping concepts. IoU scores are shown in blue.

representational entanglement varies by model, with SDv2 showing greater resilience to change in entanglement measurements post-intervention. This disparity further validates our framework's ability to detect architectural differences in bias representation that remain invisible to traditional demographic parity metrics.

Particularly striking in Table 6 is the dramatic increase in racial entanglement for certain **professions** in SDv2 compared to SD 1.5. The *firefighter* **profession** exhibits **race** -concept entanglement of 0.552 in SDv2 versus 0.353 in SD 1.5, despite identical distributional bias (RD = 0.980). Similar patterns emerge for *engineer* (0.539 vs. 0.396) and *officer* (0.540 vs. 0.347). This systematic increase suggests that architectural changes in SDv2 may have inadvertently strengthened internal coupling between racial attributes and certain **profession** concepts, despite no deterioration in distributional metrics—a finding that would remain undetected without our spatial entanglement analysis.

The relationship between initial bias severity and mitigation potential observed in SD 1.5 persists in SDv2, further confirming the model-agnostic nature of this phenomenon. **Profession** with extreme initial racial bias like *nurse* (RD = 1.000) show substantial improvements in RD (to 0.620) but negligible changes in mIoU (0.404 to 0.400), reinforcing our thesis that distributional interventions fail to address deeper representational entanglement regardless of model architecture. Conversely, **professions** with lower initial bias like *artist* show minimal changes in both metrics, suggesting a consistent pattern of intervention effectiveness across model generations.

A compelling finding emerges when comparing entanglement patterns across demographic dimensions within SDv2. While certain **professions** exhibit high entanglement for both gender and **race** (*musician*: 0.481/0.496), oth-

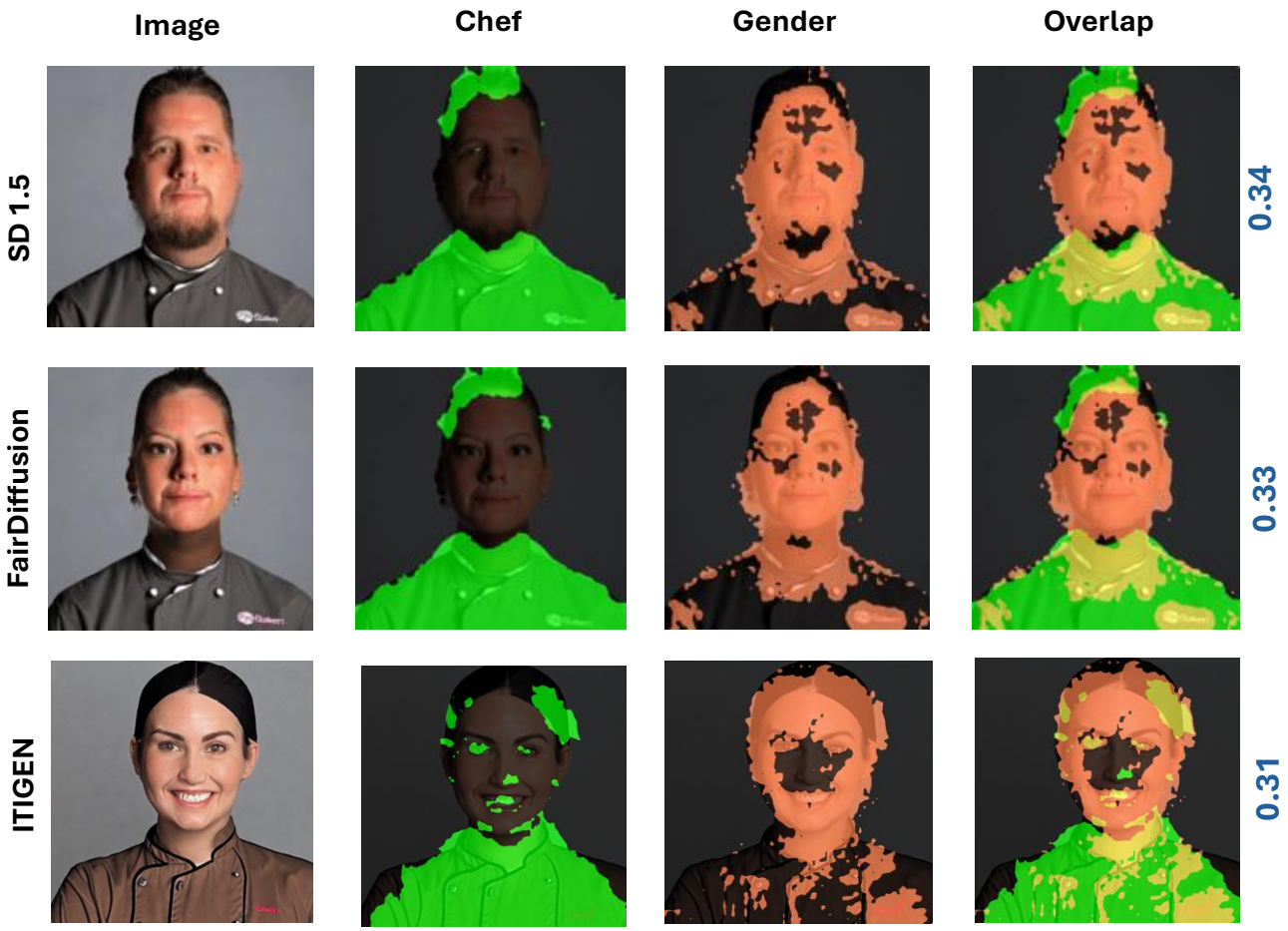

Figure 7. An overview of spatial concept overlap using a generated image of a chef. Heatmaps highlighted show the effects of attributes: Chef, Race and Overlapping concepts. IoU scores are show in blue.

ers show pronounced asymmetry (*mechanic*: 0.264/0.466). This asymmetric pattern persists across model generations but manifests differently—*mechanic* showed the lowest racial entanglement in SD 1.5 (0.175) but moderate values in SDv2 (0.466), while maintaining relatively low gender entanglement in both models. This architectural variation in demographic-specific entanglement validates our framework's sensitivity to nuanced differences in how models internally represent different types of demographic attributes.

The **profession** -specific intervention sensitivities observed in SD 1.5 reappear in SDv2 with remarkable consistency. FairDiffusion again demonstrates superior performance for *nurse* gender bias (RD = 0.020) and *pilot* racial bias (RD = 0.020), suggesting that certain **profession** concepts interact with demographic representations in consistent ways regardless of model architecture. This cross-model consistency in intervention effectiveness patterns further validates BiasMap as a generalizable framework for bias discovery and characterization across diverse text-to-image models, capable of capturing both architectural invariants and model-specific entanglement signatures.

D. Qualitative Results

We also performed qualitative analysis on individual generations. The results are shown in Figure 5, 6, and 7. SDv2 results are shown in 8 and 9.

E. Does mIoU capture conceptual entanglement?

To validate mIoU as a measure of conceptual entanglement, we examine the correlation between cross-attention map

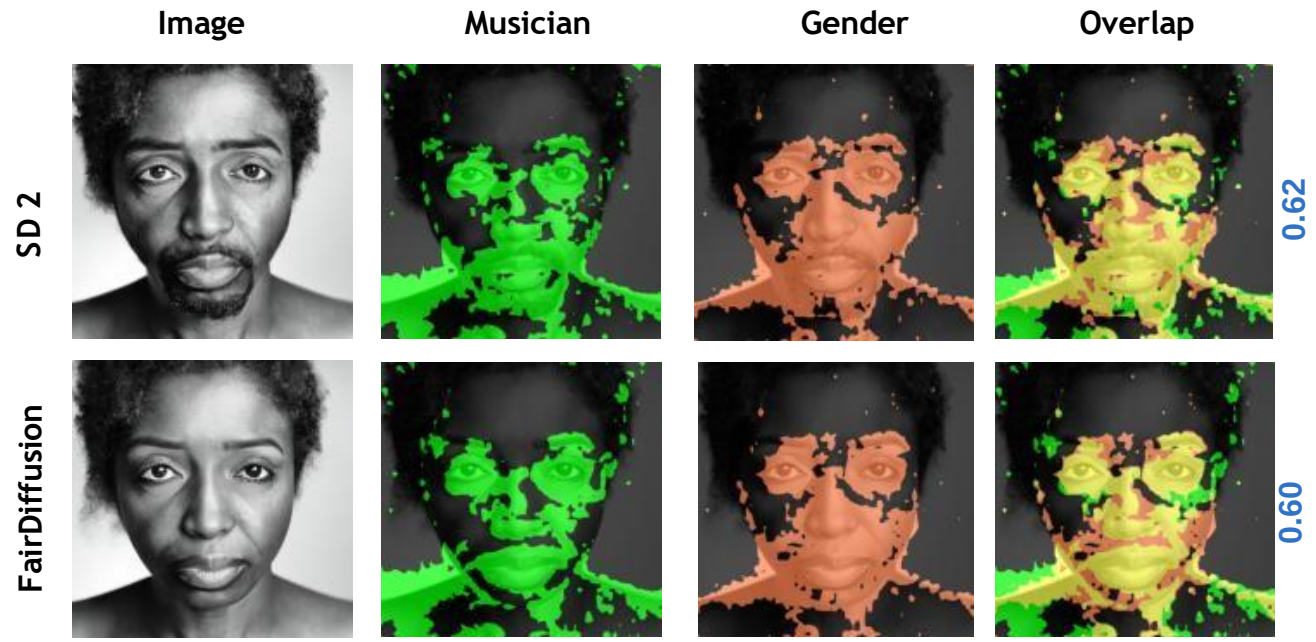

Figure 8. An overview of spatial concept overlap using a generated image of a musician, for SD2 and FD generations. Heatmaps highlighted show the effects of attributes: Musician, Race and Overlapping concepts. IoU scores are shown in blue.

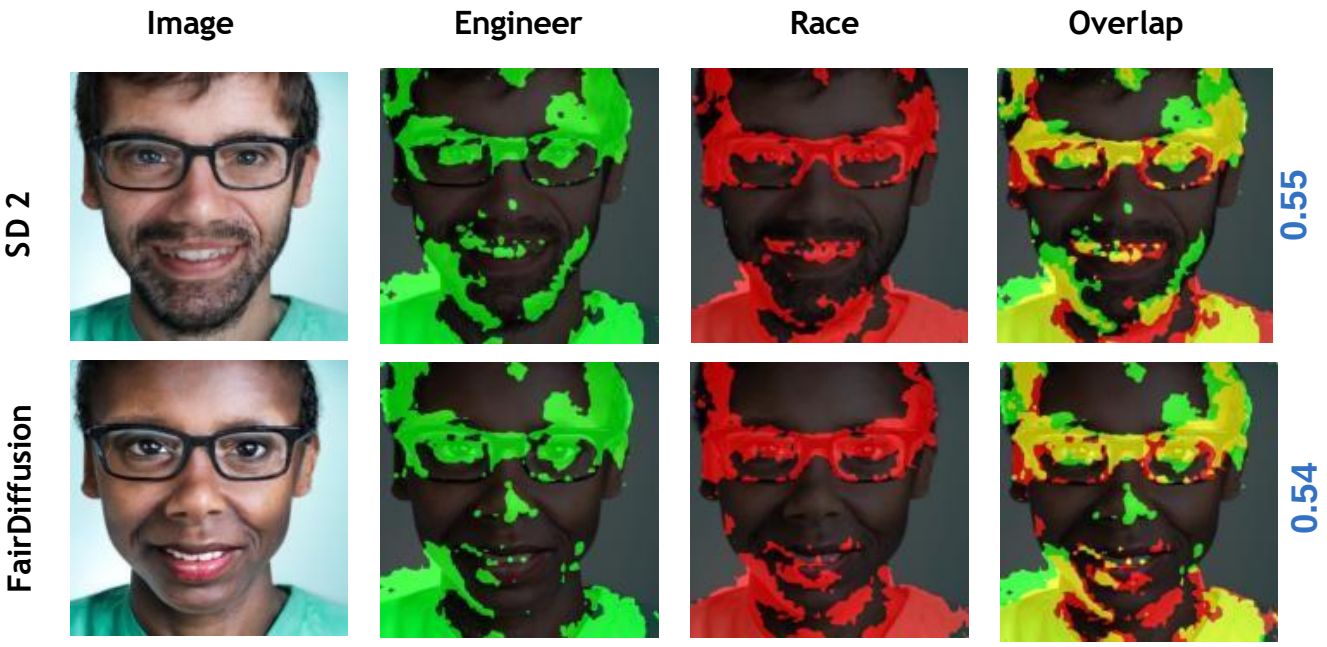

Figure 9. An overview of spatial concept overlap using a generated image of a engineer, for SD2 and FD generations. Heatmaps highlighted show the effects of attributes: Engineer, Race and Overlapping concepts. IoU scores are shown in blue.

overlap and semantic proximity in embedding space. Figure 10 presents this relationship for concepts related to **race** .We carefully select eight comparison concepts spanning a semantic gradient from closely related ("ethnicity") to distant ("quantum"). For each concept, we compute cosine similarity with the **race** anchor using a CLIP text encoder

Table 5. SDv2 gender -based Metrics with FairDiffusion

profession	mIoU ↓		RD ↓	
	SDv2	FD	SDv2	FD
firefighter	0.367	0.366	0.980	0.260
teacher	0.427	0.426	0.900	0.100
nurse	0.422	0.421	0.600	**0.020**
engineer	0.449	0.447	0.900	0.220
doctor	0.395	0.391	0.920	0.320
chef	0.312	0.310	0.900	0.300
officer	0.289	0.289	0.760	0.080
pilot	0.327	0.321	0.980	0.400
architect	0.460	0.465	0.940	0.120
lawyer	0.382	0.382	0.900	0.240
scientist	0.469	0.469	0.980	0.480
journalist	0.448	0.449	0.980	0.240
artist	**0.245**	**0.246**	0.860	0.140
musician	0.481	0.486	0.480	0.120
athlete	0.524	0.523	**0.140**	0.120
cashier	0.297	0.298	0.920	0.180
mechanic	0.264	0.265	0.960	0.120
driver	0.258	0.254	0.900	0.200
worker	0.353	0.352	0.480	0.080
waiter	0.294	0.290	1.000	0.440

Table 6. SDv2 **race** -based Metrics with FairDiffusion

profession	mIoU ↓		RD ↓	
	SDv2	FD	SDv2	FD
firefighter	0.552	0.545	0.980	0.040
teacher	0.435	0.425	0.440	0.160
nurse	0.404	0.400	1.000	0.620
engineer	0.539	0.533	0.800	0.200
doctor	0.438	0.428	0.220	0.140
chef	0.462	0.454	0.780	0.200
officer	0.540	0.524	0.880	0.080
pilot	0.540	0.555	0.560	**0.020**
architect	0.472	0.467	0.700	0.160
lawyer	0.419	0.411	0.420	0.240
scientist	0.458	0.454	0.560	0.220
journalist	0.466	0.463	**0.020**	0.100
artist	**0.340**	**0.335**	0.280	0.060
musician	0.496	0.490	0.240	0.120
athlete	0.511	0.513	0.320	**0.020**
cashier	0.347	0.334	0.540	0.100
mechanic	0.466	0.476	0.880	0.100
driver	0.499	0.479	0.680	**0.020**
worker	0.414	0.412	0.520	0.100
waiter	0.429	0.428	1.000	0.120

and measure attention map overlap using mIoU with the 70th percentile threshold. The results demonstrate that semantically similar concepts consistently exhibit higher spatial overlap in attention maps. "Ethnicity" shows both the highest semantic similarity (0.76) and the highest mIoU (0.40), while conceptually distant terms like "language" display minimal overlap (mIoU = 0.25). Intermediate concepts ("nationality," "skin," "religion") form a cluster with moderate similarity scores (0.67-0.71) and correspondingly moderate mIoU values (0.33-0.34). This monotonic relationship confirms that mIoU captures meaningful conceptual relationships rather than arbitrary correlations. The validation establishes that cross-attention maps spatially localize concepts in a manner reflecting semantic relationships, mIoU serves as a reliable proxy for conceptual association strength, and attention-based measurements capture substantive semantic associations encoded within the model's generative process. This enables confident application of mIoU for measuring conceptual entanglement between demographic attributes and **professions** in subsequent analyses, ensuring that our bias measurements reflect meaningful associations rather than measurement artifacts.

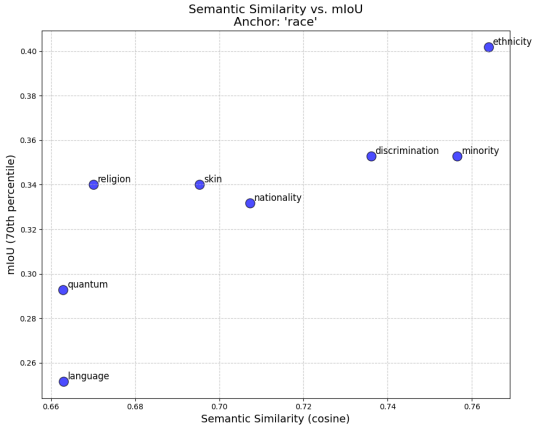

Figure 10. Correlation between semantic similarity (cosine distance in embedding space) and mean Intersection over Union (mIoU) for concepts related to "**race** ". The positive trend validates mIoU as a meaningful measure of semantic alignment in cross-attention maps.

F. Faithfulness of mIoU

To determine the optimal threshold for binarizing attention maps, we conduct a systematic analysis of mask accuracy across different threshold percentiles, as shown in Figure 11. When evaluating gender -concept attribution maps in the SD1.5 model, we observe that gender mask accuracy remains consistently high ($> 98\%$) for thresholds between the 10th and 70th percentiles, peaking at 99.67% at the 30th percentile. However, accuracy declines precipitously beyond the 70th percentile, dropping to 86% at the 90th percentile. Conversely, the complement mask accuracy increases steadily with higher thresholds, reaching

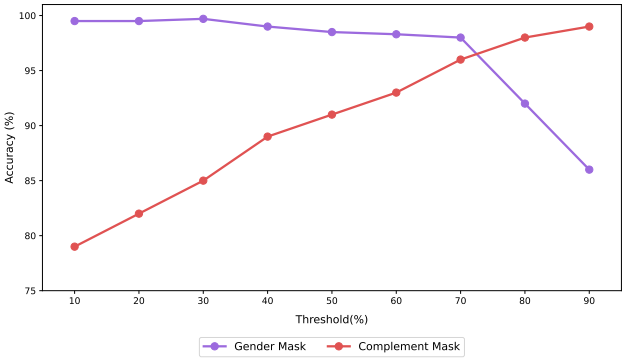

Figure 11. Mask accuracy analysis for gender attribution across different thresholds. The gender mask maintains high accuracy ($\geq 98\%$) up to the 70th percentile threshold, after which it declines rapidly. The complement mask shows steadily increasing accuracy with higher thresholds. The intersection point at approximately threshold 75 provides empirical justification for our selection of the 70th percentile threshold.

peak performance (99.06%) at the 90th percentile.

The intersection point of these trends occurs at approximately the 75th percentile, suggesting an optimal balance between gender attribution and its complement. Based on this analysis, we selected the 70th percentile as our threshold for all subsequent experiments, representing the highest threshold value before gender mask accuracy begins to deteriorate significantly. This threshold ensures robust attribution of demographic concepts while maintaining discriminative power between related concepts.