# CLASSIFIER-TO-BIAS: Toward Unsupervised Automatic Bias Detection for Visual Classifiers

Supplementary Material

## Table of Contents


Appendix A: Limitations of our approach . . . 2
Appendix B: Domain shift and retrieval error . . . 3
Appendix C: Additional quantitative results with varying thresholds . . . 4
Appendix D: Additional qualitative analyses based on C2B bias scores . . . 6
Appendix E: Evaluation of the LLM-proposed biases . . . 9
Appendix F: VQA-based evaluation of the retrieval system . . . 10
Appendix G: Details and comparisons of VQA models . . . 11
Appendix H: Retrieved images diversity . . . 12
Appendix I: Prompts used for LLM bias proposal . . . 13
Appendix J: Comparison between different LLMs for bias proposal . . . 15
Appendix K: Embedding-based bias matching details and examples . . . 17
Appendix L: Ground-truth bias matrices visualizations . . . 19

## A. Limitations of our approach

As C2B is the first bias detection method in a truly unsupervised setting, it comes with some limitations. We explore these limitations below, focusing on the two core steps of C2B: bias proposal and image retrieval.

**LLM-based bias proposal.** Large language models may have limited knowledge, carry their own biases [21, 50], and are known to be prone to hallucinations [5, 31, 67]. They may propose irrelevant biases or miss biases that could only be found from the data. This limitation stems from our top-down approach, which relies on explicit proposals rather than discovering biases purely from observed failures. In contrast, bottom-up approaches [14, 33, 34, 71] that mine biases from annotated model failures are themselves constrained by the available data and annotations: a model may exhibit biases not represented in the dataset. These two paradigms can be seen as complementary. Additionally, LLM proposals are sensitive to prompting; while we report the prompts used in Appendix I, alternative prompting strategies or fine-tuned LLMs could yield different or improved results. The modularity of our framework allows for future integration of more specialized or domain-adapted LLMs to improve proposal coverage.

**Image retrieval.** The accuracy of C2B strongly depends on the quality of the retrieval system. If retrieved images fail to match the intended target and bias attributes, bias scores will be unreliable. As shown in Sec. 4.4 and Appendix F, current retrieval systems leave room for improvement: across tasks and retrieval sources, less than 50% of retrieved images correspond to both the intended target and bias classes according to VQA evaluation. This difficulty is due in part to the compositional complexity of the retrieval queries, which vision-language models often struggle with [26, 62]. Using larger-scale datasets such as DataComp-1B [20] or LAION-5B [55], and improved embedding models like SigLIP [66], could enhance retrieval precision and compositional understanding. Furthermore, in domain-specific contexts (such as medical imaging), biases may not be directly visible. However, C2B's LLM-based proposal mechanism can surface such biases: for instance, suggesting biases related to hospital type or imaging device. A natural future extension could involve metadata-based retrieval, leveraging contextual attributes beyond visual cues to detect subtle domain-specific biases.

**Ethical statement and broader impact.** This work aims to contribute to fairer and more transparent AI by enabling unsupervised detection of biases in pre-trained classifiers. We conduct this research responsibly and with attention to ethical considerations. Nonetheless, due to practical constraints, some socially sensitive attributes (*e.g.*, gender or ethnicity) are treated as closed sets for research purposes only. In addition, C2B inherently reflects the limitations and potential biases of the LLMs and retrieval systems it relies on, and thus may not detect all possible biases. Our intention is not to discriminate against any social group, but rather to raise awareness of the challenges involved in bias discovery and to promote responsible use and auditing of AI models.

## B. Domain shift and retrieval error

C2B relies on images retrieved from large-scale external datasets or web sources (*e.g.*, CC12M or Bing). A potential limitation is that these images may differ in style, composition, or distribution compared to the classifier's training domain (CelebA or ImageNet), introducing domain shift and retrieval noise. This can affect bias scoring accuracy. To better understand the impact of domain shift and retrieval errors, we conducted additional experiments under more controlled conditions. We considered two alternative retrieval strategies:

• Retrieval from domain-specific datasets (CELEBA and IMAGENET): We perform CLIP-based image retrieval directly from the evaluation datasets (CelebA or ImageNet-X) instead of external sources. This reduces domain shift but does not remove retrieval noise, as we still rely on CLIP similarity rather than annotations.

• Retrieval from labeled subsets (CELEBA-GT and IMAGENET-GT): We further constrain retrieval to subsets of the datasets where the target class matches the desired target label. This eliminates ambiguity in the target class, though bias attributes are still unknown and must be inferred via CLIP. This setup reduces some noise and simulates a scenario where partial information is available, making it equivalent to the open-set bias detection setting of B2T.

Table 5. Proportion (%) of ground-truth biases detected on CelebA (GT → Detected) and of detected biases corresponding to ground-truth ones (Detected → GT). FH=False Hit. Agreement between detected biases and VQA on CelebA.

| METHOD | GT → DETECTED | | | DETECTED → GT | | | VQA |
|---|---|---|---|---|---|---|---|
| | HIT(↑) | FH(↓) | MISS (↓) | HIT(↑) | FH (↓) | MISS(↓) | AGREEMENT |
| | FaceXFormer | | | | | | |
| C2B (CELEBA) | 11.67 | 7.43 | 80.90 | 12.57 | 10.02 | **77.41** | 0.27 |
| C2B (CELEBA-GT) | 11.43 | 7.50 | 81.07 | 12.24 | 9.00 | 78.77 | **0.32** |

Table 6. Proportion (%) of ground-truth biases detected on ImageNet-X (GT → Detected) and of detected biases corresponding to ground-truth ones (Detected → GT). FH=False Hit. Agreement between detected biases and VQA on CelebA.

| METHOD | GT → DETECTED | | | DETECTED → GT | | | VQA |
|---|---|---|---|---|---|---|---|
| | HIT(↑) | FH(↓) | MISS (↓) | HIT(↑) | FH (↓) | MISS(↓) | AGREEMENT |
| | ResNet50_V2 | | | | | | |
| C2B (IMAGENET) | 8.25 | 9.61 | 82.13 | 2.73 | 2.86 | 94.41 | 0.28 |
| C2B (IMAGENET-GT) | 5.24 | 4.48 | 90.28 | **3.23** | 2.54 | 94.24 | **0.40** |
| | ResNet101_V2 | | | | | | |
| C2B (IMAGENET) | 8.98 | 9.56 | 81.46 | 2.64 | 2.80 | 94.56 | 0.27 |
| C2B (IMAGENET-GT) | 4.09 | 4.28 | 91.63 | **3.40** | 3.02 | **93.58** | **0.44** |
| | ResNet152_V2 | | | | | | |
| C2B (IMAGENET) | 8.71 | 9.29 | 81.99 | 2.71 | 2.85 | 94.44 | 0.28 |
| C2B (IMAGENET-GT) | 3.39 | 3.49 | 93.11 | **3.62** | 2.70 | **93.68** | **0.47** |
| | ViT_B_16_SWAG | | | | | | |
| C2B (IMAGENET) | 8.61 | 9.50 | 81.89 | 2.48 | 2.76 | 94.76 | 0.29 |
| C2B (IMAGENET-GT) | 4.07 | 3.81 | 92.12 | **3.30** | 3.19 | **93.51** | **0.48** |

Interestingly, retrieving from the ground-truth dataset itself does not lead to significantly more stable or accurate detection of annotated biases. Bias scores remain comparable to those obtained from external retrieval sources, suggesting that domain shift is not the only source of error. Retrieval noise and the difficulty of capturing subtle bias attributes remain major factors.

Retrieving from labeled subsets results in a trade-off: the hit rate for ground-truth biases decreases, but the false hit rate also reduces, indicating greater precision. Importantly, VQA-based evaluation reveals that both controlled retrieval methods (especially retrieval from labeled subsets) achieve higher agreement with VQA-labeled biases. This suggests that, although C2B detects fewer ground-truth biases in these settings, the detected biases are semantically more meaningful and visually verifiable. These findings highlight that retrieval quality and dataset alignment both impact C2B's performance, and they confirm that part of the observed noise originates from imperfect retrieval. Improving retrieval precision could therefore further enhance bias detection.

## C. Additional quantitative results with varying thresholds

In Sec. 4.2.1, we chose to present results with a similarity threshold of 0.9 for embedding-based bias matching, and a bias detection threshold of 0.05 for all methods. While we believe that these thresholds represent the best trade-off we could find to detect and match similar biases while avoiding false positives, we present additional results with different threshold values in this section for completeness.

Table 7. Proportion (%) of ground-truth biases detected on CelebA (GT → Detected) and of detected biases corresponding to ground-truth ones (Detected → GT) **with a similarity threshold of 0.8**. FH=False Hit.

| Method | GT → Detected | | | Detected → GT | | |
|---|---|---|---|---|---|---|
| | Hit(↑) | FH(↓) | Miss (↓) | Hit(↑) | FH (↓) | Miss(↓) |
| | FaceXFormer | | | | | |
| B2T [33] | 8.48 | **6.52** | 85.00 | 12.25 | **8.78** | 78.98 |
| C2B (Bing) | **22.69** | 17.26 | **60.05** | **27.30** | 18.21 | **54.49** |
| C2B (cc12m) | 19.61 | 17.67 | 62.72 | 23.22 | 20.53 | 56.25 |
| C2B (CelebA) | 21.06 | 16.79 | 62.15 | 22.75 | 20.73 | 56.44 |

Table 8. Proportion (%) of ground-truth biases detected on CelebA (GT → Detected) and of detected biases corresponding to ground-truth ones (Detected → GT) **with a similarity threshold of 0.95**. FH=False Hit.

| Method | GT → Detected | | | Detected → GT | | |
|---|---|---|---|---|---|---|
| | Hit(↑) | FH(↓) | Miss (↓) | Hit(↑) | FH (↓) | Miss(↓) |
| | FaceXFormer | | | | | |
| B2T [33] | 2.27 | **0.78** | 96.96 | 3.31 | **1.14** | 95.55 |
| C2B (Bing) | **6.16** | 2.91 | 90.93 | **7.82** | 2.70 | 89.48 |
| C2B (cc12m) | 5.20 | 3.69 | 91.10 | 6.88 | 3.71 | 89.41 |
| C2B (CelebA) | 6.08 | 3.05 | **90.88** | 6.40 | 5.13 | **88.46** |

Table 9. Proportion (%) of ground-truth biases detected on ImageNet-X (GT → Detected) and of detected biases corresponding to ground-truth ones (Detected → GT) **with a similarity threshold of 0.8**. FH=False Hit.

| Method | GT → Detected | | | Detected → GT | | |
|---|---|---|---|---|---|---|
| | Hit(↑) | FH(↓) | Miss (↓) | Hit(↑) | FH (↓) | Miss(↓) |
| | ResNet50_V2 | | | | | |
| B2T [33] | 11.34 | **10.21** | 78.45 | 3.68 | **3.25** | 93.07 |
| C2B (Bing) | 22.50 | 25.17 | 52.33 | 7.68 | 8.38 | **83.93** |
| C2B (cc12m) | **28.06** | 29.28 | **42.65** | **7.78** | 7.85 | 84.38 |
| C2B (ImageNet) | 22.96 | 24.79 | 52.25 | 7.76 | 7.89 | 84.35 |
| | ResNet101_V2 | | | | | |
| B2T [33] | 10.96 | **9.69** | 79.34 | 3.50 | **3.18** | 93.32 |
| C2B (Bing) | 22.07 | 24.43 | 53.50 | 7.41 | 8.22 | **84.37** |
| C2B (cc12m) | **27.59** | 29.62 | **42.79** | 7.32 | 7.93 | 84.75 |
| C2B (ImageNet) | 23.69 | 24.77 | 51.54 | **7.50** | 7.98 | 84.53 |
| | ResNet152_V2 | | | | | |
| B2T [33] | 10.91 | **10.21** | 78.87 | 3.49 | **3.38** | 93.13 |
| C2B (Bing) | 21.45 | 23.68 | 54.86 | 7.37 | 8.36 | **84.27** |
| C2B (cc12m) | **27.59** | 29.81 | **42.60** | 7.28 | 7.91 | 84.81 |
| C2B (ImageNet) | 22.86 | 24.62 | 52.53 | **7.68** | 7.80 | 84.51 |
| | ViT_B_16_SWAG | | | | | |
| B2T [33] | 10.10 | **10.19** | 79.71 | 3.23 | **3.33** | 93.44 |
| C2B (Bing) | 20.63 | 22.42 | 56.95 | 6.70 | 7.52 | 85.78 |
| C2B (cc12m) | **27.23** | 29.92 | **42.85** | 6.90 | 7.43 | **85.67** |
| C2B (ImageNet) | 23.21 | 24.80 | 52.00 | **6.91** | 7.17 | 85.91 |

Table 10. Proportion (%) of ground-truth biases detected on ImageNet-X (GT → Detected) and of detected biases corresponding to ground-truth ones (Detected → GT) **with a similarity threshold of 0.95**. FH=False Hit.

| Method | GT → Detected | | | Detected → GT | | |
|---|---|---|---|---|---|---|
| | Hit(↑) | FH(↓) | Miss (↓) | Hit(↑) | FH (↓) | Miss(↓) |
| | ResNet50_V2 | | | | | |
| B2T [33] | 0.19 | **0.50** | 99.30 | 0.06 | **0.15** | 99.79 |
| C2B (Bing) | 2.77 | 2.40 | 94.83 | 0.94 | 0.76 | 98.30 |
| C2B (cc12m) | **3.96** | 3.49 | **92.55** | **1.07** | 0.95 | **97.98** |
| C2B (ImageNet) | 2.71 | 2.69 | 94.60 | 0.92 | 0.84 | 98.24 |
| | ResNet101_V2 | | | | | |
| B2T [33] | 0.37 | **0.54** | 99.09 | 0.12 | **0.13** | 99.75 |
| C2B (Bing) | 2.93 | 2.49 | 94.58 | 0.96 | 0.79 | 98.25 |
| C2B (cc12m) | **3.96** | 4.00 | **92.03** | **0.98** | 1.05 | **97.97** |
| C2B (ImageNet) | 3.09 | 2.99 | 93.93 | 0.96 | 0.91 | 98.14 |
| | ResNet152_V2 | | | | | |
| B2T [33] | 0.34 | **0.33** | 99.33 | 0.11 | **0.08** | 99.81 |
| C2B (Bing) | 3.06 | 2.45 | 94.49 | **1.10** | 0.86 | **98.04** |
| C2B (cc12m) | **3.81** | 4.10 | **92.10** | 0.94 | 1.00 | 98.06 |
| C2B (ImageNet) | 3.02 | 3.00 | 93.98 | 0.93 | 0.94 | 98.13 |
| | ViT_B_16_SWAG | | | | | |
| B2T [33] | 0.32 | **0.26** | 99.42 | 0.10 | **0.08** | 99.82 |
| C2B (Bing) | 2.84 | 2.35 | 94.81 | 0.91 | 0.78 | 98.31 |
| C2B (cc12m) | **3.84** | 3.78 | **92.39** | **0.94** | 0.90 | **98.16** |
| C2B (ImageNet) | 3.06 | 3.06 | 93.88 | 0.80 | 0.89 | 98.30 |

Table 11. Proportion (%) of ground-truth biases detected on CelebA (GT → Detected) and of detected biases corresponding to ground-truth ones (Detected → GT) **with a bias detection threshold of 0.01**. FH=False Hit.

| Method | GT → Detected | | | Detected → GT | | |
|---|---|---|---|---|---|---|
| | Hit(↑) | FH(↓) | Miss (↓) | Hit(↑) | FH (↓) | Miss(↓) |
| | FaceXFormer | | | | | |
| B2T [33] | 6.01 | **3.46** | 90.53 | 10.72 | **6.07** | 83.20 |
| C2B (Bing) | **13.26** | 9.48 | **77.26** | **16.03** | 11.91 | 72.06 |
| C2B (cc12m) | 12.29 | 9.44 | 78.28 | 15.83 | 12.77 | **71.40** |
| C2B (CelebA) | 13.07 | 8.76 | 78.17 | 15.50 | 11.84 | 72.65 |

Table 12. Proportion (%) of ground-truth biases detected on CelebA (GT → Detected) and of detected biases corresponding to ground-truth ones (Detected → GT) **with a bias detection threshold of 0.1**. FH=False Hit.

| Method | GT → Detected | | | Detected → GT | | |
|---|---|---|---|---|---|---|
| | Hit(↑) | FH(↓) | Miss (↓) | Hit(↑) | FH (↓) | Miss(↓) |
| | FaceXFormer | | | | | |
| B2T [33] | 4.29 | **0.95** | 94.76 | 5.51 | **1.37** | 93.12 |
| C2B (Bing) | 9.78 | 3.94 | 86.28 | 9.07 | 4.68 | 86.25 |
| C2B (cc12m) | 9.46 | 6.89 | 83.65 | 11.23 | 4.72 | 84.04 |
| C2B (CelebA) | **12.73** | 4.27 | **83.00** | **11.99** | 6.42 | **81.58** |

First, we present results with a different similarity threshold for embedding-based bias matching. In Tab. 7, we present results on CelebA with a similarity threshold of 0.8. In Tab. 8, we present results on CelebA with a similarity threshold

Table 13. Proportion (%) of ground-truth biases detected on ImageNet-X (GT → Detected) and of detected biases corresponding to ground-truth ones (Detected → GT) **with a bias detection threshold of 0.01**. FH=False Hit.

| Method | GT → Detected | | | Detected → GT | | |
|---|---|---|---|---|---|---|
| | Hit(↑) | FH(↓) | Miss (↓) | Hit(↑) | FH (↓) | Miss(↓) |
| | ResNet50_V2 | | | | | |
| B2T [33] | 2.33 | **2.19** | 95.47 | 0.92 | **0.80** | 98.28 |
| C2B (Bing) | 11.51 | 12.09 | 76.40 | 3.23 | 3.53 | 93.24 |
| C2B (cc12m) | **14.32** | 15.02 | **70.66** | **3.49** | 3.52 | **92.99** |
| C2B (ImageNet) | 12.05 | 12.06 | 75.89 | 3.44 | 3.26 | 93.30 |
| | ResNet101_V2 | | | | | |
| B2T [33] | 2.58 | **2.06** | 95.35 | 1.01 | **0.80** | 98.19 |
| C2B (Bing) | 10.67 | 11.80 | 77.53 | 3.10 | 3.29 | 93.61 |
| C2B (cc12m) | **14.18** | 14.80 | **71.03** | **3.27** | 3.53 | **93.20** |
| C2B (ImageNet) | 12.02 | 12.31 | 75.67 | 3.20 | 3.31 | 93.49 |
| | ResNet152_V2 | | | | | |
| B2T [33] | 2.53 | **1.81** | 95.66 | 0.99 | **0.72** | 98.29 |
| C2B (Bing) | 10.71 | 11.51 | 77.78 | 3.13 | 3.51 | 93.36 |
| C2B (cc12m) | **14.13** | 15.09 | **70.78** | **3.37** | 3.61 | **93.02** |
| C2B (ImageNet) | 11.79 | 12.45 | 75.77 | 3.24 | 3.47 | 93.29 |
| | ViT_B_16_SWAG | | | | | |
| B2T [33] | 2.22 | **2.04** | 95.75 | 0.95 | **0.86** | 98.19 |
| C2B (Bing) | 10.10 | 10.38 | 79.52 | 3.07 | 3.09 | 93.84 |
| C2B (cc12m) | **14.32** | 14.42 | **71.27** | **3.36** | 3.40 | **93.24** |
| C2B (ImageNet) | 12.02 | 12.31 | 75.68 | 3.20 | 3.30 | 93.50 |

Table 14. Proportion (%) of ground-truth biases detected on ImageNet-X (GT → Detected) and of detected biases corresponding to ground-truth ones (Detected → GT) **with a bias detection threshold of 0.1**. FH=False Hit.

| Method | GT → Detected | | | Detected → GT | | |
|---|---|---|---|---|---|---|
| | Hit(↑) | FH(↓) | Miss (↓) | Hit(↑) | FH (↓) | Miss(↓) |
| | ResNet50_V2 | | | | | |
| B2T [33] | 2.49 | **2.08** | 95.43 | 0.61 | **0.51** | 98.88 |
| C2B (Bing) | 3.50 | 4.58 | 91.93 | 1.74 | 2.30 | 95.96 |
| C2B (cc12m) | **7.06** | 6.52 | **86.41** | **2.07** | 2.11 | **95.81** |
| C2B (ImageNet) | 4.69 | 5.34 | 89.97 | 1.93 | 2.03 | 96.03 |
| | ResNet101_V2 | | | | | |
| B2T [33] | 2.93 | **1.79** | 95.29 | 0.66 | **0.48** | 98.85 |
| C2B (Bing) | 3.88 | 4.26 | 91.86 | 1.89 | 1.96 | 96.15 |
| C2B (cc12m) | **6.15** | 7.46 | **86.39** | **2.13** | 2.32 | **95.56** |
| C2B (ImageNet) | 5.13 | 5.11 | 89.76 | **2.13** | 2.02 | 95.85 |
| | ResNet152_V2 | | | | | |
| B2T [33] | 2.68 | **1.49** | 95.84 | 0.64 | **0.38** | 98.98 |
| C2B (Bing) | 3.78 | 4.36 | 91.86 | 1.77 | 2.13 | 96.10 |
| C2B (cc12m) | **6.09** | 7.18 | **86.72** | **1.96** | 2.31 | **95.73** |
| C2B (ImageNet) | 4.82 | 5.28 | 89.90 | 1.92 | 1.98 | 96.10 |
| | ViT_B_16_SWAG | | | | | |
| B2T [33] | 2.16 | **1.95** | 95.89 | 0.56 | **0.49** | 98.95 |
| C2B (Bing) | 3.66 | 3.58 | 92.76 | 1.89 | 1.63 | 96.48 |
| C2B (cc12m) | **6.20** | 6.60 | **87.20** | **1.99** | 1.98 | **96.02** |
| C2B (ImageNet) | 4.83 | 5.58 | 89.59 | 1.70 | 1.96 | 96.34 |

of 0.95. In Tab. 9, we present results on ImageNet-X with a similarity threshold of 0.8. In Tab. 10, we present results on ImageNet-X with a similarity threshold of 0.95.

Second, we present results with a bias detection threshold. In Tab. 11, we present results on CelebA with a bias detection threshold of 0.01. In Tab. 12, we present results on CelebA with a bias detection threshold of 0.1. In Tab. 13, we present results on ImageNet-X with a bias detection threshold of 0.01. In Tab. 14, we present results on ImageNet-X with a bias detection threshold of 0.1.

Our results show that C2B's advantage over B2T is consistent across a wide range of threshold values. Lower thresholds increase recall but also lead to more false positives, whereas stricter thresholds reduce false hits at the expense of missing subtle biases. Across all configurations, C2B maintains a higher proportion of ground-truth biases detected, confirming the method's robustness. In addition, C2B consistently exhibits a lower miss rate than B2T, despite operating in a fully unsupervised setting. These results highlight that our approach is stable and reliable, and that performance is not overly sensitive to hyperparameter choices.

## D. Additional qualitative analyses based on C2B bias scores

In this section, we propose additional qualitative analyses based on the bias scores assigned by C2B.

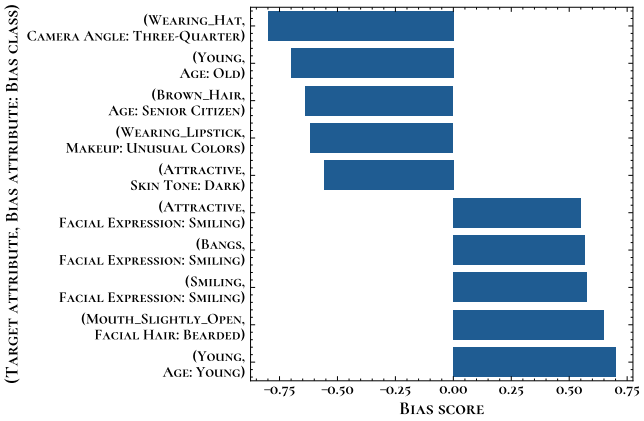


Figure 7. Strongest detected biases of **FaceXFormer** over all target attributes on **face attribute classification**.

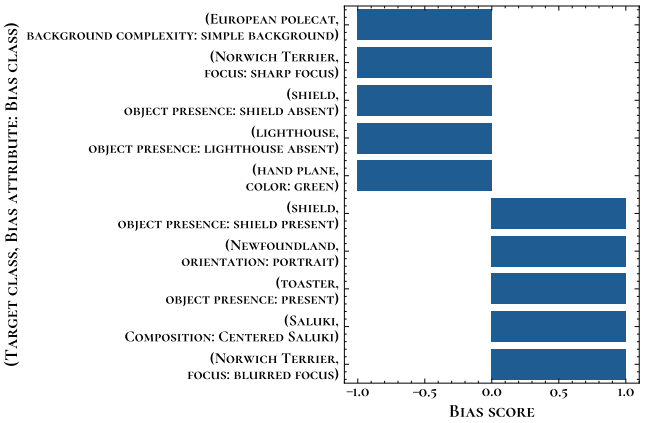


Figure 8. Strongest detected biases of **ResNet50_V2** over all target classes on **image classification.**

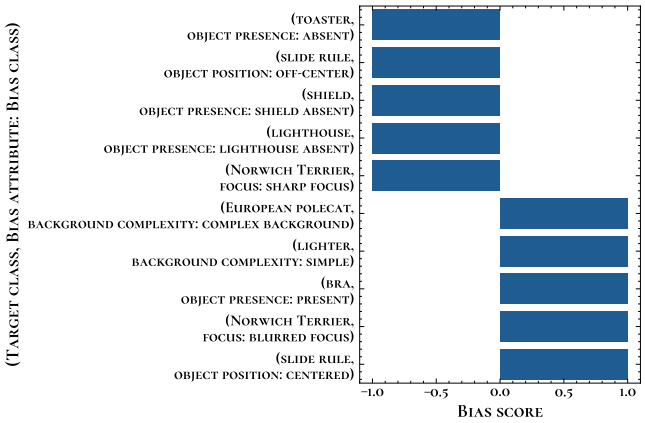


Figure 9. Strongest detected biases of **ResNet101_V2** over all target classes on **image classification.**

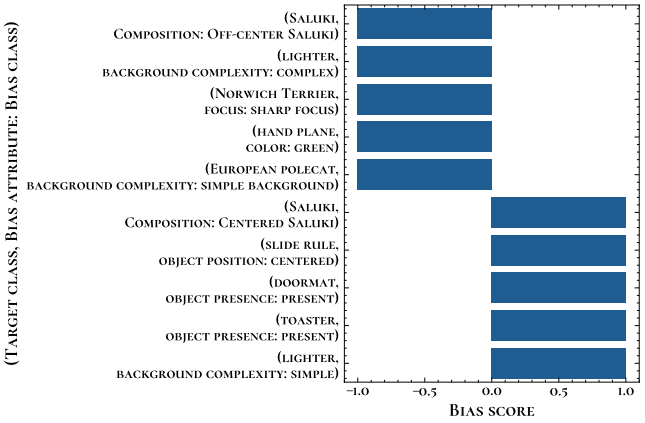


Figure 10. Strongest detected biases of **ResNet152_V2** over all target classes on **image classification.**

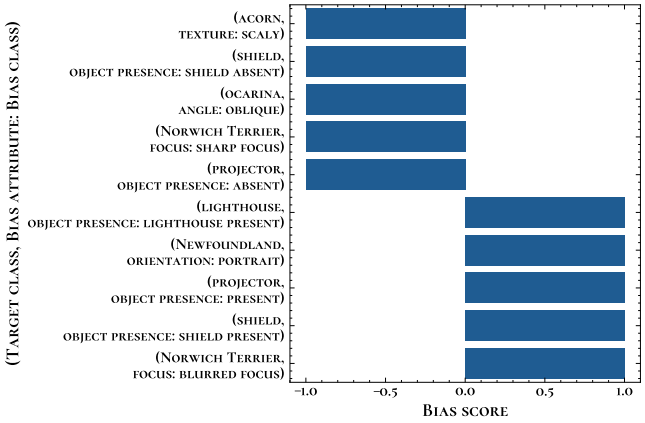


Figure 11. Strongest detected biases of **ViT_B_16_SWAG** over all target classes on **image classification.**

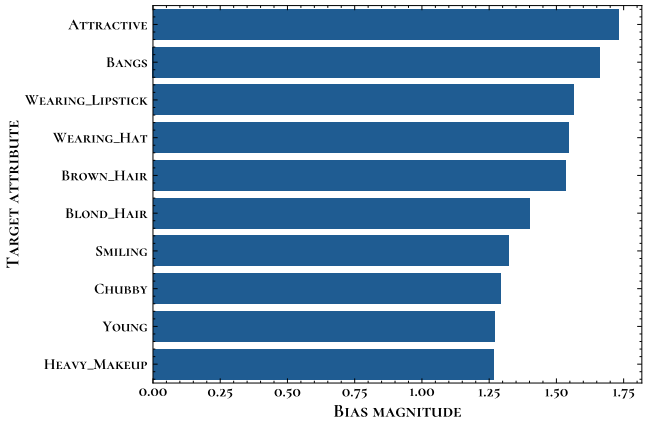


Figure 12. Most bias-affected target attributes for **FaceXFormer.**

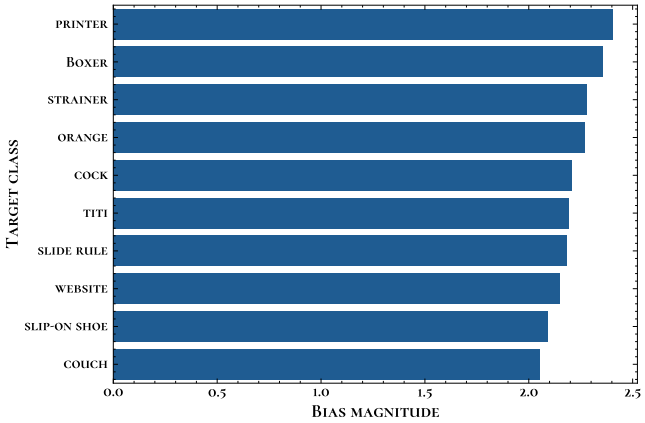


Figure 13. Most bias-affected target classes for **ResNet50_V2**.

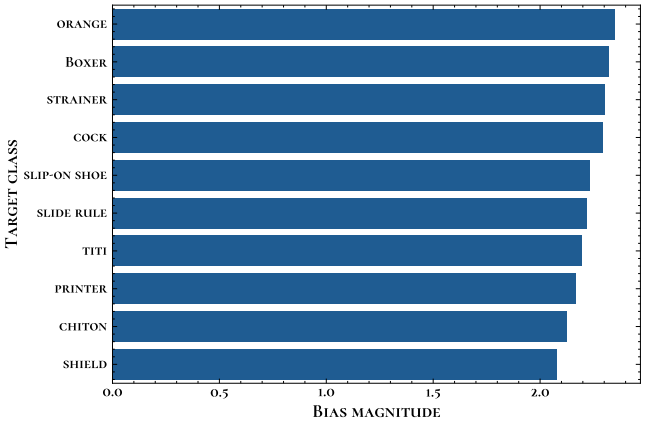


Figure 14. Most bias-affected target classes for **ResNet101_V2**.

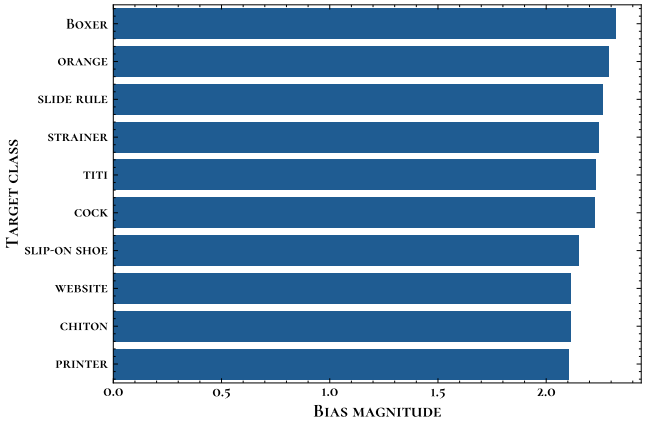


Figure 15. Most bias-affected target classes for **ResNet152_V2**.

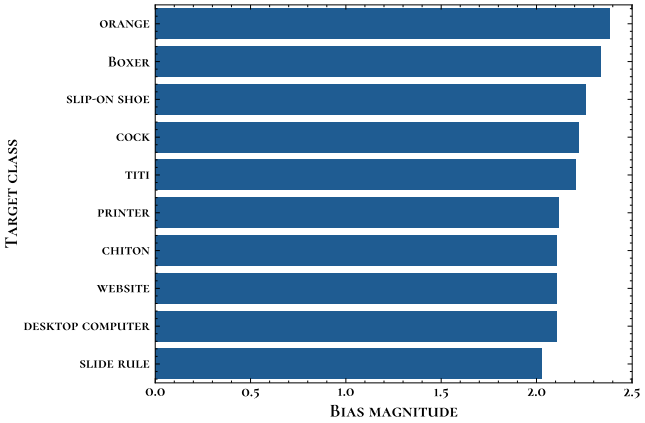


Figure 16. Most bias-affected target classes for **ViT_B_16_SWAG**.

For the first type of analysis, we show the 5 strongest positive biases and the 5 strongest negative biases of a model over all target attributes and classes for a given task. In Fig. 7, we show these biases for FaceXFormer face attribute classification. In Figs. 8 to 11, we show these biases for ResNet50_V2, ResNet101_V2, ResNet152_V2, and ViT_B_16_SWAG_E2E_V1, respectively, on image classification.

These figures can be very interesting to inform the user about the strongest biases of a model, and also allow to discover

new biases (such as "camera angle: three-quarter" for the *wearing hat* attribute in Fig. 7 or "color: green" for the *hand plane* class in Figs. 8 and 10), but they also illustrate some failure cases of C2B. We can see in Fig. 7 that "Age" was proposed as a bias attribute when classifying the *young* attribute, or that "object presence/absence" was proposed for several different classes in Figs. 8 to 11. These inevitably affect the performance of the classifier, but cannot be considered as "biases". The implementation of a bias filtering mechanism could lead to improvements in this regard.

For the second type of analysis, we propose to show the target attributes/classes that are the most affected by biases. For each target attribute/class, we define the bias magnitude as the L2 norm of the vector containing the bias scores for all bias classes. In Fig. 12, we show the most bias-affected target attributes and their bias magnitude for FaceXFormer face attribute classification. In Figs. 13 to 16, we show the most bias-affected target classes and their bias magnitude for ResNet50_V2, ResNet101_V2, ResNet152_V2, and ViT_B_16_SWAG_E2E_V1, respectively, on image classification.

These figures can be useful to inform the user about which target attributes/classes are the most affected by biases when using a specific model. For instance, we can see in Fig. 7 that the *attractive* target attribute is the most affected by biases. From this, the user may want to look at the various bias scores for the different bias attributes and classes that were proposed for the *attractive* target attribute. We can also see in Figs. 13 to 16 that *boxer* and *orange* are two of the most bias-affected classes across all four tested models.

## E. Evaluating LLM-proposed biases

C2B relies on large language models (LLMs) to propose candidate bias attributes based on a textual description of the task and target classes. While this approach enables an unsupervised and task-agnostic bias discovery process, it also raises the question of how well the LLM-proposed biases align with known ground-truth biases and whether they contribute to discovering novel biases. We analyze two aspects:

• Coverage of ground-truth biases (LLM MISS): We measure the proportion of ground-truth biases that are included among the LLM-proposed biases.

• Novel bias discovery (NEW BIAS): We evaluate what proportion of detected biases (*i.e.*, biases with high bias scores identified by C2B) are not part of the annotated ground-truth set, indicating the discovery of previously unannotated biases.

Table 15. Proportion (%) of ground-truth biases detected on CelebA (GT → Detected) and of detected biases corresponding to ground-truth ones (Detected → GT). FH=False Hit.

| METHOD | GT → DETECTED | | | | DETECTED → GT | | | |
|---|---|---|---|---|---|---|---|---|
| | HIT(↑) | FH(↓) | RETRIEVAL MISS(↓) | LLM MISS(↓) | HIT(↑) | FH (↓) | NOT A BIAS(↓) | NEW BIAS(↑) |
| | FaceXFormer | | | | | | | |
| C2B (BING) | **12.29** | 6.88 | **9.94** | 70.89 | **14.18** | 7.34 | **9.58** | **68.91** |
| C2B (CC12M) | 10.76 | 7.75 | 10.60 | 70.89 | 12.75 | 8.15 | 12.08 | 67.02 |
| C2B (CELEBA) | 11.67 | 7.43 | 10.01 | 70.89 | 12.57 | 10.02 | 9.97 | 67.44 |

Table 16. Proportion (%) of ground-truth biases detected on ImageNet-X (GT → Detected) and of detected biases corresponding to ground-truth ones (Detected → GT). FH=False Hit.

| METHOD | GT → DETECTED | | | | DETECTED → GT | | | |
|---|---|---|---|---|---|---|---|---|
| | HIT(↑) | FH(↓) | RETRIEVAL MISS(↓) | LLM MISS(↓) | HIT(↑) | FH (↓) | NOT A BIAS(↓) | NEW BIAS(↑) |
| | ResNet50_V2 | | | | | | | |
| C2B (BING) | 7.80 | 8.53 | 16.79 | 66.88 | 2.60 | 2.98 | **11.81** | **82.61** |
| C2B (CC12M) | **11.18** | 11.30 | **10.65** | 66.88 | **2.99** | 2.87 | 12.54 | 81.60 |
| C2B (IMAGENET) | 8.25 | 9.61 | 15.25 | 66.88 | 2.73 | 2.86 | 12.38 | 82.03 |
| | ResNet101_V2 | | | | | | | |
| C2B (BING) | 7.91 | 8.41 | 17.01 | 66.67 | 2.56 | 2.77 | **12.10** | **82.58** |
| C2B (CC12M) | **11.21** | 11.68 | **10.45** | 66.67 | **2.76** | 2.96 | 12.87 | 81.41 |
| C2B (IMAGENET) | 8.98 | 9.56 | 14.80 | 66.67 | 2.64 | 2.80 | 12.37 | 82.19 |
| | ResNet152_V2 | | | | | | | |
| C2B (BING) | 7.53 | 7.93 | 17.77 | 66.77 | 2.75 | 2.79 | **11.56** | **82.90** |
| C2B (CC12M) | **11.10** | 12.01 | **10.12** | 66.77 | **2.81** | 2.95 | 12.83 | 81.42 |
| C2B (IMAGENET) | 8.71 | 9.29 | 15.22 | 66.77 | 2.71 | 2.85 | 12.33 | 82.11 |
| | ViT_B_16_SWAG | | | | | | | |
| C2B (BING) | 7.72 | 7.18 | 18.03 | 67.07 | 2.48 | 2.48 | **11.71** | **83.34** |
| C2B (CC12M) | **10.85** | 11.07 | **11.02** | 67.07 | **2.64** | 2.70 | 13.20 | 81.46 |
| C2B (IMAGENET) | 8.61 | 9.50 | 14.83 | 67.07 | 2.48 | 2.76 | 12.62 | 82.15 |

As shown in Tabs. 15 and 16, the LLM proposals cover approximately 29% of the ground-truth biases on CelebA and 33% on ImageNet-X. While this partial coverage reflects the inherent limitations of relying on a language model without task-specific supervision, it also confirms that LLMs can generate relevant and meaningful bias candidates in a wide variety of domains.

At the same time, a large fraction of the biases detected by C2B are not present in the ground-truth annotations: around 68% on CelebA and 82% on ImageNet-X. This demonstrates C2B 's ability to discover novel, previously unannotated biases. Some of these newly surfaced biases are confirmed through qualitative inspection and VQA-based validation to reflect real, systematic model behaviors. These results highlight that while LLM proposals do not exhaustively cover all known biases, they provide a rich and diverse starting point for unsupervised bias discovery, enabling C2B to go beyond closed-set annotations and uncover previously overlooked spurious correlations.

## F. VQA-based evaluation of the retrieval system

As mentioned in Appendix A, C2B critically depends on the accuracy of the retrieval. In Sec. 4.4, we propose to measure the accuracy of the CLIP-based retrieval by retrieving on a labeled dataset and measuring the recall @ K. While this is the only type of possible evaluation using ground-truth annotations, this is not perfectly representative of our use case, because the captions that were used could not contain the LLM-proposed biases, but had to rely on labeled attributes. Moreover, this evaluation can only measure the accuracy of the CLIP-based retrieval, as Bing cannot be used to retrieve images from a local database.

For these reasons, we propose to evaluate the actual retrieved data with a visual question answering (VQA) model. This allows to measure the quality of the retrieved images, by asking the VQA if the attributes that we want are indeed in the images (both the target and bias classes). This also allows a direct comparison between Bing and the CLIP-based retrieval (both on CC12M and the evaluation dataset itself, as seen in Appendix B). We present our results in Tab. 17 for the face attribute classification task, and in Tab. 18 for the image classification task.

Table 17. Accuracy of the retrieval according to the VQA on the face attribute classification task.

| Retrieval method | Accuracy | | |
|---|---|---|---|
| | Target | Bias | Both |
| Bing | **78.33** | 60.33 | 46.49 |
| CLIP + CC12M | 67.72 | 63.69 | 42.43 |
| CLIP + CelebA | 72.98 | **65.37** | **46.65** |

Table 18. Accuracy of the retrieval according to the VQA on the image classification task.

| Retrieval method | Accuracy | | |
|---|---|---|---|
| | Target | Bias | Both |
| Bing | **89.10** | 48.63 | **42.82** |
| CLIP + CC12M | 77.66 | **50.29** | 38.05 |
| CLIP + ImageNet | 86.86 | 47.31 | 40.35 |

In both Tabs. 17 and 18, we can see that Bing-retrieved images are generally more likely to contain the right target class, while CLIP-retrieved images are more likely to contain the right bias class. According to the VQA, on average, Bing is more accurate than CLIP-retrieval. However, as discussed in Appendix A, there is still a lot of room for improvement, as the accuracy for "both" (the case where the image contains the right target class and the right bias class) is below 50% for all methods, according to the VQA. For additional details about the VQA model, please refer to Appendix G, where the accuracy of the VQA itself is measured.

## G. Details and comparisons of VQA models

Table 19. VQA models chosen for comparison.

| MODEL | VERSION | PARAMS | SIZE (GB) | RELEASED |
|---|---|---|---|---|
| LLAVA-1.5 | `1.5-13b-hf` | 13B | 26.69 | 10/2023 |
| LLAVA-NEXT | `v1.6-vicuna-13bf-hf` | 13B | 26.69 | 01/2024 |

For the results presented in Sec. 4.2.2, we chose to pseudo-label the images with LLaVA-1.5-13B [38, 39], which we found to be significantly faster than LLaVA-NeXT [40], with comparable accuracy for our use case (see Tabs. 19 and 20).

To label images with C2B-proposed biases, we use multiple-choice questions about bias attributes, with answer choices representing the proposed bias classes. In the case of B2T, we ask binary yes-no questions about the presence of keywords associated to bias in images.

Table 20. Comparing VQA models across various tasks and datasets. ACC. is the accuracy, BM the informedness metric, and TIME the run time per image (in ms).

| | CELEBA FACE ATTRIBUTES | | | | | IMAGENET CLASSES | | IMAGENET-X FACTORS | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| MODEL | ACC. | TPR | TNR | BM | TIME | ACC. | TIME | ACC. | TPR | TNR | BM | TIME |
| LLaVA-1.5 | 76.50 | **75.18** | 76.88 | **0.521** | **120** | **85.29** | **105** | 39.09 | **85.15** | 30.28 | 0.154 | **101** |
| LLaVA-NeXT | **78.80** | 66.57 | **82.29** | 0.489 | 217 | 71.20 | 433 | **48.33** | 72.20 | **43.76** | **0.160** | 592 |

In Tab. 20, we compare the performance of LLaVA-1.5 and LLaVA-NeXT (both 13B versions) to predict binary face attributes on Celeba (left), target classes on ImageNet (center), and binary factors on ImageNet-X (right).

Because of class imbalance, we choose to also show true positive rate (TPR) and true negative rate (TNR) when classifying binary face attributes on CelebA or binary factors on ImageNet-X. These metrics are combined into (bookmaker) informedness (BM), also known as Youden's J statistic, defined as $BM = TPR + TNR - 1$. Informedness is proportional to balanced accuracy and is considered to be a more appropriate metric to assess the random guessing level of a classifier [13].

Overall, we found LLaVA-1.5 and LLaVA-NeXT to have comparable performance on the two datasets, but noted that LLaVA-1.5 had a tendency to reply more positively than LLaVA-NeXT (especially visible when looking at TPR and TNR). Moreover, we observed that, in our case, LLaVA-1.5 was better at following instructions (giving an answer within the given choices) than LLaVA-NeXT, with LLaVA-NeXT being more subject to hallucinations.

## H. Retrieved images diversity

The quality of retrieved images plays a crucial role in C2B 's performance. While retrieval accuracy is important, the diversity of retrieved samples for each caption is equally critical. Low diversity can lead to overly homogeneous image sets that do not accurately capture the variability of the intended bias attribute, potentially skewing bias scores or reducing the robustness of bias detection.

To quantify retrieval diversity, we compute the average pairwise CLIP embedding similarity between all the retrieved images for a given target class (TARGET CLASS SIM.), as well as for each caption (BIAS CLASS SIM.). A higher average similarity indicates less diversity (more homogeneous images), while a lower value indicates more varied visual content. We compare diversity scores for images retrieved from all retrieval sources, including domain-specific datasets, to assess which source provides more diverse image sets.

Table 21. Average cosine similarity between retrieved image embeddings across both tasks.

| | FACE ATTRIBUTE CLASSIFICATION | | | IMAGE CLASSIFICATION | | |
|---|---|---|---|---|---|---|
| RETRIEVAL SOURCE | BING | CC12M | CELEBA | BING | CC12M | IMAGENET |
| TARGET CLASS SIM. | **0.60** | 0.65 | 0.66 | **0.75** | **0.75** | 0.78 |
| BIAS CLASS SIM. | **0.65** | 0.72 | 0.71 | **0.77** | 0.78 | 0.80 |

As shown in Tab. 21, Bing retrieval produces the most diverse image sets, followed by CC12M, followed by domain-specific datasets (CelebA and ImageNet). This confirms that web-scale retrieval sources provide richer visual variety. We hypothesize that this higher diversity contributes to better detection of complex or subtle biases and may explain why Bing-based retrieval often yields slightly better bias discovery performance in our experiments. These findings suggest that diversity-aware filtering or weighting strategies could be beneficial future improvements for bias assessment pipelines like C2B.

# I. Prompts used for LLM bias proposal

In the following, we describe the textual elements of C2B: the user input, the bias generation prompts, and the caption ones.

**User input.** The only input needed from the user is a textual description of the task, including the target classes. We provide the descriptions that we used for our two example classification tasks in Figs. 17 and 18. New task descriptions can easily be added to detect biases for different classification tasks.

```
In the domain of face attribute classification, models are trained to identify
attributes from images of human faces. These attributes can include, but are not
limited to, age, gender, expression, presence of accessories (e.g., glasses, hats),
facial hair, skin tone, and any distinctive features.
```

Figure 17. Task description for face attribute classification.

```
In the domain of image classification, models are trained to assign a label or category
to an entire image, identifying objects, scenes, or concepts based on visual content.
```

Figure 18. Task description for image classification.

**Bias generation prompts.** As explained in Sec. 3.3, different biases are generated for each target class. The LLM is prompted in a "chat completion" mode using the ChatML format [46]. First, a *system* prompt (Fig. 19) is given with general instructions about the task at hand. Then, a *user* prompt is given, containing information about the task (name and description), the target attribute, the target class, as well as some additional instructions (Fig. 20) for additional guidance. The expected response format is also provided to the LLM (*JSON* Schema).

```
You are a helpful intelligent assistant knowledgeable about computer vision. Our
objective is to identify potential visual biases in pre-trained computer vision
classifier models. Given the description of a computer vision task, the name of a
target attribute, and the name of the class that we are trying to identify, generate
a list of potential visual biases that a classifier may have. These biases must be
identifiable in images and should reflect common issues that can arise from the
training data or model architecture.

Instructions:
1. Generate a list of visually-identifiable bias attributes that could influence the
performance of a pre-trained classifier for a given task.
2. For each bias attribute, provide a list of bias classes that represent all the
potential values of this attribute.
3. Output the list of bias attributes and their classes in JSON format.
```

Figure 19. System prompt for bias generation.

```
Think about what characteristic or feature of the image could impact the performance
of the model. Think about potential spurious correlations and potential failure types.
```

Figure 20. Additional instructions given in the user prompt for bias generation.

**Caption template prompts.** As explained in Sec. 3.4, the first step is to generate a caption template for the task, which will then be adapted by the LLM to fit different combinations of target and bias classes. As is done for bias generation, a system prompt with general instructions is first given (Fig. 21), and a user prompt is given in a second phase, containing the task name, the task description, and additional instructions (Fig. 22).

```
You are a large language model designed to generate simple captions for images related
to a specific computer vision task. Given the description of a computer vision task,
your task is to generate a simple caption template that would work for any image
relevant to this task. This template will be used at the beginning of every caption, so
keep it general. Think about general characteristics that apply to all images of this
task. The template should only correspond to the beginning of a sentence, and end with
"{}", a placeholder that will be personalized in the future to generate all captions for
this task.
```

Figure 21. System prompt for caption template generation.

```
Think about the type of image (e.g., a photo) and what it should contain. Keep the
template simple and general. Do not explicitly mention the task in the template. Avoid
unnecessary words and do not make sentences.
```

Figure 22. Additional instructions given in the user prompt for caption template generation.

**Caption generation prompts.** Finally, captions are generated for each combination of target and bias classes. In practice, the LLM is only prompted for each bias attribute (for each target class), and produces the captions for all the corresponding bias classes at once. This reduces the number of calls to the LLM, allows the captions to be more consistent across bias classes, and helps the generation process by providing more context to the LLM. Again, a system prompt (Fig. 23) is initially given to the LLM, before a user prompt containing information about the task (name and description), the target class, the bias attribute and bias classes, the caption template, as well as additional instructions (Fig. 24) is also given.

```
You are a large language model designed to generate simple captions for images related
to a specific computer vision task. Your objective is to generate captions as simple
combinations of attributes to retrieve images that match the task as well as the
attributes. Given the description of a computer vision task, a target class, a bias
attribute, and bias classes for this attribute, your task is to generate simple captions
that describe images that combine the given elements. To help generate relevant captions,
a template is also provided, you can adapt the template in any way you want, even change
it if necessary. The captions should make grammatical and logical sense, be relevant to
the task, and always combine the given target class and bias class. Please avoid using
negations, provide the opposite meaning of the negated word.
```

Figure 23. System prompt for caption generation.

```
Do not introduce any new bias in the captions. Do not add new attributes. The only
attributes included in the caption should be the given target class and the given
bias class.
```

Figure 24. Additional instructions given in the user prompt for caption generation.

## J. Comparison between different LLMs for bias proposal

To generate potential biases, we have chosen to compare several lightweight quantized versions of recent state-of-the-art LLMs of comparable sizes, *i.e.*, Gemma [45], Llama [18], and Phi [1]. General specifications are shown in Tab. 22.

Table 22. Lightweight LLMs chosen for comparison.

| MODEL | VERSION | PARAMS | QUANTIZATION | SIZE (GB) | RELEASED |
|---|---|---|---|---|---|
| GEMMA | `2-9b-it` | 9B | `Q6_K_L` | 7.81 | 06/2024 |
| LLAMA | `3.1-8B-Instruct` | 8B | `Q8_0` | 8.54 | 07/2024 |
| PHI | `3-medium-128k-instruct` | 14B | `Q4_K_M` | 8.57 | 05/2024 |

Table 23. Comparison between lightweight LLMs for bias proposal. $|\mathcal{B}|$ is the average number of bias attributes per class, $|\mathcal{B}|$ the average number of bias classes per class, and TIME is the average run time in seconds.

| TASK | MODEL | $\lvert\mathcal{B}\rvert$ | $\lvert\mathbf{B}\rvert$ | $\lvert\mathbf{B}\rvert/\lvert\mathcal{B}\rvert$ | TIME | TIME/$\lvert\mathbf{B}\rvert$ |
|---|---|---|---|---|---|---|
| CELEBA | GEMMA | 7.58 | 25.68 | **3.39** | 13.76 | 0.54 |
| | LLAMA | **11.00** | **35.45** | 3.26 | 11.60 | 0.33 |
| | PHI | 6.28 | 20.90 | 3.28 | **4.11** | **0.20** |
| IMAGENET | GEMMA | 6.16 | 23.17 | **3.76** | 12.74 | 0.55 |
| | LLAMA | **9.74** | **28.72** | 2.95 | 8.72 | 0.30 |
| | PHI | 6.34 | 18.62 | 2.94 | **3.73** | **0.20** |

In Table 23, we show the average number of proposed bias attributes ($|\mathcal{B}|$) and corresponding bias classes per target class ($|\mathbf{B}|$), as well as the execution time per target class (TIME), for both face attribute classification and image classification. With our prompting strategy, Llama was the LLM giving us the largest number of potential biases, while Phi was surprisingly fast. On average, Gemma proposes more bias classes per bias attribute, but is slower than both Llama and Phi.

Qualitatively, proposed biases are fairly similar between the three tested LLMs. We provide some examples of bias that were proposed by Gemma, LLama, and Phi in Figs. 25 to 27, respectively, for the *smiling* target attribute of CelebA. On average, we have found the biases proposed by Llama to be slightly more relevant than the ones proposed by Gemma and Phi, which drove our decision to choose it for C2B.

```
[{"bias␣attribute": "Lighting",
  "bias␣classes": ["Bright", "Dim", "Shadowed"]},
 {"bias␣attribute": "Pose",
  "bias␣classes": ["Front-facing", "Profile", "Three-quarter"]},
 {"bias␣attribute": "Facial␣Expression␣Context",
  "bias␣classes": ["Happy", "Sad", "Neutral", "Angry", "Surprised"]},
 {"bias␣attribute": "Image␣Quality",
  "bias␣classes": ["High␣Resolution", "Low␣Resolution", "Blurry", "Distorted"]},
 {"bias␣attribute": "Cultural␣Background",
  "bias␣classes": ["Western", "Eastern", "African", "Other"]},
 {"bias␣attribute": "Age",
  "bias␣classes": ["Young", "Adult", "Elderly"]}]
```

Figure 25. Example of biases proposed by Gemma, for the *smiling* target attribute of CelebA.

```
[{"bias␣attribute": "Lighting",
  "bias␣classes": ["Bright", "Dim", "Shaded"]},
 {"bias␣attribute": "Facial␣Expression",
  "bias␣classes": ["Smiling", "Neutral", "Frowning"]},
 {"bias␣attribute": "Glasses",
  "bias␣classes": ["Present", "Absent"]},
 {"bias␣attribute": "Hats␣and␣Headwear",
  "bias␣classes": ["Present", "Absent"]},
 {"bias␣attribute": "Facial␣Hair",
  "bias␣classes": ["Present", "Absent"]},
 {"bias␣attribute": "Skin␣Tone",
  "bias␣classes": ["Fair", "Medium", "Dark"]},
 {"bias␣attribute": "Age",
  "bias␣classes": ["Young", "Old"]},
 {"bias␣attribute": "Image␣Quality",
  "bias␣classes": ["High␣Resolution", "Low␣Resolution"]},
 {"bias␣attribute": "Camera␣Angle",
  "bias␣classes": ["Frontal", "Profile"]},
 {"bias␣attribute": "Background␣Clutter",
  "bias␣classes": ["Clean", "Cluttered"]}]
```


Figure 26. Example of biases proposed by Llama, for the *smiling* target attribute of CelebA.


```
[{"bias␣attribute": "Facial␣Expression",
  "bias␣classes": ["Smiling", "Neutral", "Frowning"]},
 {"bias␣attribute": "Lighting␣Conditions",
  "bias␣classes": ["Bright␣Light", "Dim␣Light", "Backlight", "Shadowed␣Face"]},
 {"bias␣attribute": "Facial␣Accessories",
  "bias␣classes": ["Glasses", "Hats", "Masks", "None"]},
 {"bias␣attribute": "Skin␣Tone",
  "bias␣classes": ["Light␣Skin", "Dark␣Skin", "Tanned␣Skin", "Pale␣Skin"]},
 {"bias␣attribute": "Facial␣Hair",
  "bias␣classes": ["Beard", "Mustache", "Clean-Shaven", "None"]},
 {"bias␣attribute": "Age␣Group",
  "bias␣classes": ["Young", "Middle␣Age", "Older␣Adults"]}]
```


Figure 27. Example of biases proposed by Phi, for the *smiling* target attribute of CelebA.

## K. Embedding-based bias matching details and examples

The proposed biases may be names that do not match those in the ground truth (*e.g.*, even due to synonyms, such as "male" and "man"). Thus, for the evaluation based on ground-truth annotations presented in Sec. 4.2.1, we match proposed and ground-truth biases using cosine similarity of their respective SBERT embeddings.

In practice, we found similarity scores based on embeddings of single words to be unreliable. The embeddings were actually computed on captions containing the bias attributes/classes, *i.e.*, "A photo of a young person", instead of "Young".

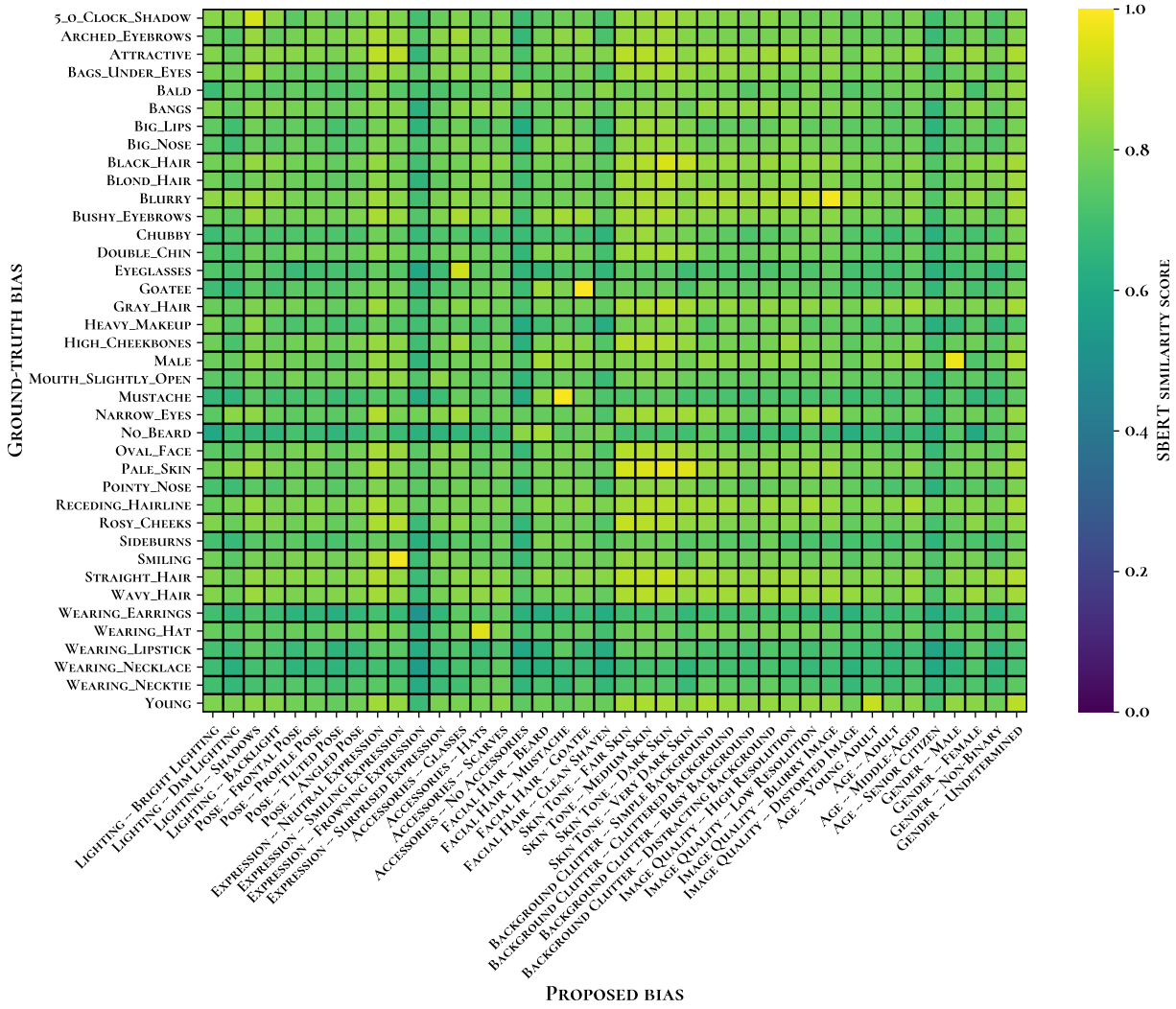


Figure 28. Example of SBERT similarity scores between ground-truth attributes (used as ground-truth biases) and biases proposed by C2B for the *brown hair* attribute on CelebA.

In Fig. 28, we show an example of similarity scores between ground-truth attributes (used as ground-truth biases) and bias proposed by C2B for the *brown hair* attribute on CelebA. First, a similarity threshold is defined as a minimum score for two biases to be matched. This threshold was set the 0.9 for the results we present in Sec. 4.2.1. Matching biases is then an iterative process, where the most similar pair is matched and removed from the similarity matrix, until no pairs above the similarity threshold are left.

In Tab. 24, we show examples of similarity scores between ground-truth attributes (used as ground-truth biases) and bias proposed by C2B for the *attractive* attribute on CelebA. We can see that a similarity score of 0.95 would have missed a true match, while a similarity score of 0.8 would have been too low and would have resulted in too many false matches.

Table 24. Examples of similarity scores (SIM. SCORE) between ground-truth attributes (used as ground-truth biases) and bias proposed by C2B for the *attractive* attribute on CelebA.

| GROUND-TRUTH | PROPOSED | | SIM. |
|---|---|---|---|
| | ATTRIBUTE | CLASS | SCORE |
| Smiling | Facial Expression | Smiling | 0.98 |
| Heavy_Makeup | Makeup | Heavy | 0.98 |
| Young | Age | Young | 0.98 |
| Eyeglasses | Glasses | Present | 0.93 |
| Pale_Skin | Skin Tone | Dark | 0.88 |
| Blurry | Lighting | Dim | 0.84 |
| Rosy_Cheeks | Lighting | Bright | 0.84 |
| Goatee | Facial Hair | Present | 0.83 |
| 5_o_Clock_Shadow | Lighting | Shaded | 0.81 |

We show additional quantitative results varying the similarity threshold in Appendix C (Tabs. 7 to 10).

## L. Ground-truth bias matrices visualizations

We follow the definition of bias given in Sec. 3.1 to compute ground-truth biases, which are then used for the evaluation based on ground-truth annotations presented in Sec. 4.2.1.

In this section, we show the ground-truth bias matrices for FaceXFormer on CelebA, as well as ResNet50_V2, ResNet101_V2, ResNet152_V2, and ViT_B_16_SWAG_E2E_V1 on ImageNet-X.

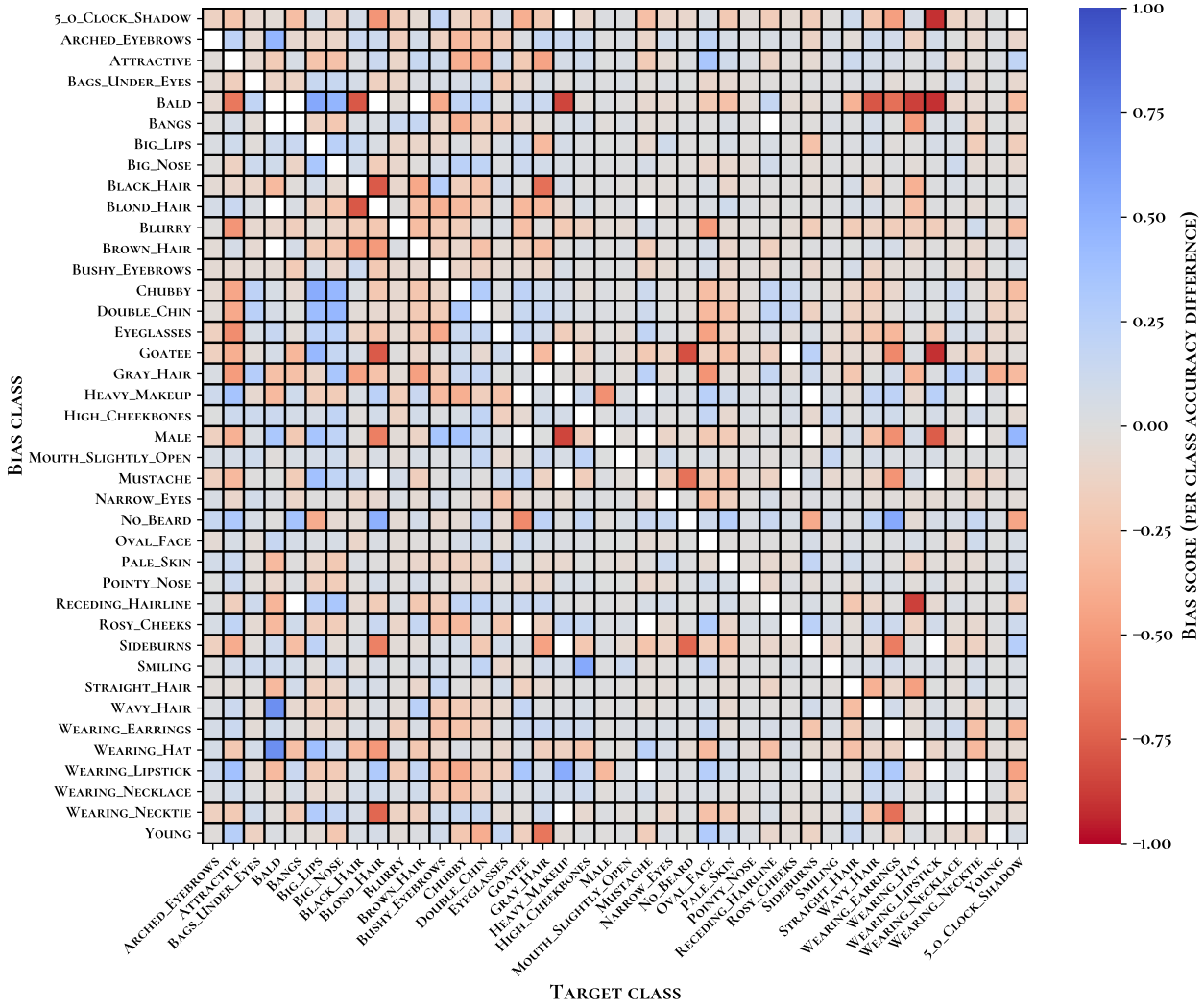


Figure 29. Ground-truth bias matrix of FaceXFormer on CelebA.

In Fig. 29, we show the ground-truth bias matrix of FaceXFormer for all classes on CelebA, where per the class accuracy is equivalent to the true positive rate for each binary attribute.

For ImageNet-X, as it would be impossible to see individual biases in full ground-truth bias matrices over all 1000 classes without zooming in, we provide bias matrices for the top 50 classes with the strongest biases for each model in Figs. 30 to 33.

For all ground-truth bias matrices, a positive bias score (blue color) indicates a higher accuracy when the bias is present, and a negative bias score (red color) indicates a lower accuracy when bias is present. A white square indicates that the bias could not be measured because there was no example in the dataset.

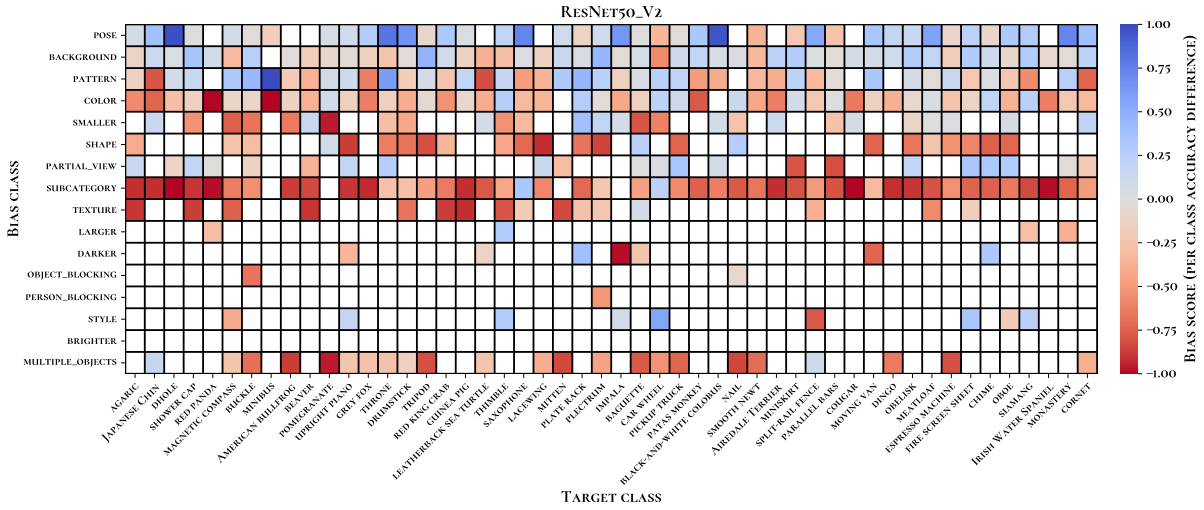


Figure 30. Ground-truth bias matrix of ResNet50_V2 for the 50 classes with the strongest biases on ImageNet-X.

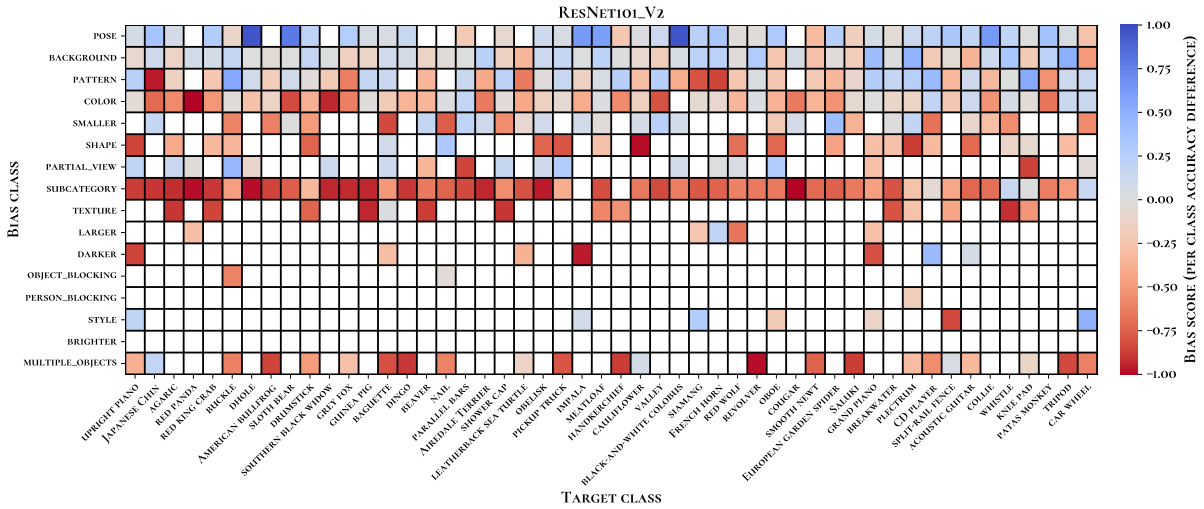


Figure 31. Ground-truth bias matrix of ResNet101_V2 for the 50 classes with the strongest biases on ImageNet-X.

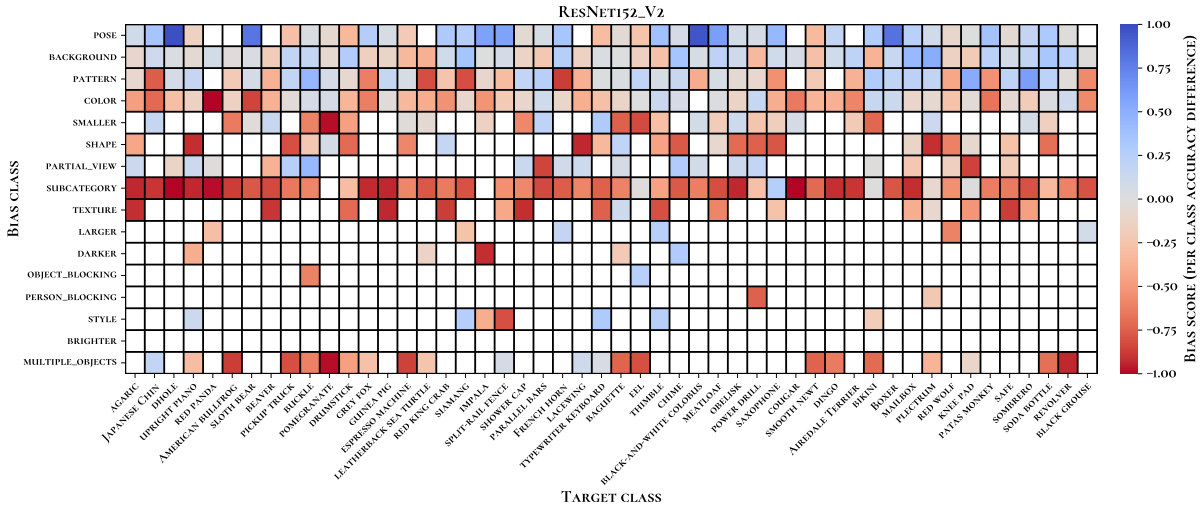


Figure 32. Ground-truth bias matrix of ResNet152_V2 for the 50 classes with the strongest biases on ImageNet-X.

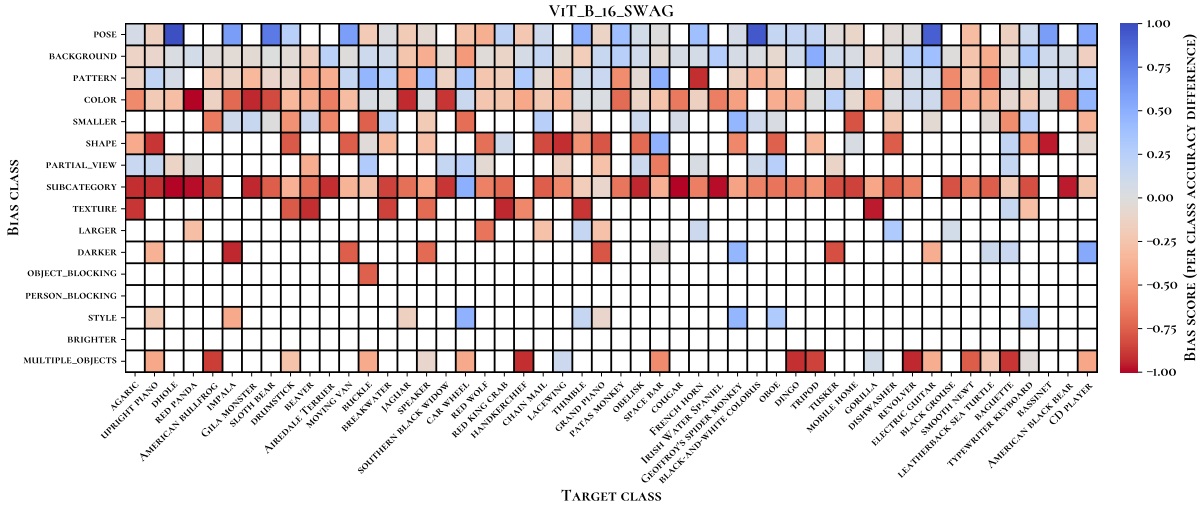


Figure 33. Ground-truth bias matrix of ViT_B_16_SWAG_E2E_V1 for the 50 classes with the strongest biases on ImageNet-X.