

**DEVELOPING METRICS FOR DETECTING GENDER BIAS
IN IMAGE DATASETS USING CONTEXTUAL FACTORS:
OBJECTS, SCENES, AND SPATIAL RELATIONSHIPS**

E.M.Amandi Mithila Ekanayake

IT21387562

BSc (Hons) degree in Information Technology Specializing in Data
Science

Department of Information Technology

Sri Lanka Institute of Information Technology

April 2025

**DEVELOPING METRICS FOR DETECTING GENDER BIAS
IN IMAGE DATASETS USING CONTEXTUAL FACTORS:
OBJECTS, SCENES, AND SPATIAL RELATIONSHIPS**

E.M.Amandi Mithila Ekanayake

(IT21387562)

Dissertation submitted in partial fulfillment of the requirements for the
Bachelor of BSc (Hons) degree in Information Technology Specializing in
Data Science

Department of Information Technology


Sri Lanka Institute of Information Technology

April 2025

DECLARATION

I declare that this is my own work and this dissertation¹ does not incorporate without acknowledgement any material previously submitted for a Degree or Diploma in any other University or institute of higher learning and to the best of my knowledge and belief it does not contain any material previously published or written by another person except where the acknowledgement is made in the text.

Also, I hereby grant to Sri Lanka Institute of Information Technology, the nonexclusive right to reproduce and distribute my dissertation, in whole or in part in print, electronic or other medium. I retain the right to use this content in whole or part in future works (such as articles or books).

Signature: 

Date: 04/11/2025

The above candidate has carried out research for the bachelor's degree Dissertation under my supervision.

Signature of the Supervisor:

Date: 04/11/2025

ABSTRACT

Contextual bias in AI-based image classification refers to the unintended influence of surrounding visual elements such as objects, spatial positioning, and scene semantics on classification outcomes. While existing fairness metrics often focus on demographic or co-occurrence imbalances, they fail to capture deeper contextual dependencies in visual data. This research proposes the **Unified Bias Metric (UBM)**, a novel, interpretable framework for detecting gender bias in image datasets by quantifying both object-level and scene-level contextual influences. The metric integrates two components: **Object Influence Score (OIS)**, which accounts for relative object size, 3D distance, and depth; and the **Scene Similarity Bias (SSB)**, which evaluates alignment with gendered scene embeddings using models like CLIP and Places365. A total of **nine experimental approaches**, including PCA, SHAP, regression models, and ensemble learning, were implemented across **852 gender-annotated COCO images**. These images feature gender-stereotyped objects (e.g., handbags, sports balls) and are analyzed under varying bias scenarios. Results demonstrate that contextual elements significantly influence gender classification, even when the subject remains constant. By providing object-wise and context-aware bias scores, the UBM offers a scalable, statistically validated, and explainable diagnostic tool. This work lays the foundation for developing open-source toolkits for fairness auditing in visual AI systems.

Keywords—Gender Bias, Contextual Fairness, Scene Similarity Bias, Object Influence Score, Unified Bias Metric, Explainable AI, Image Classification Ethics

ACKNOWLEDGEMENT

First and foremost, I would like to express my deepest gratitude to my supervisor, Dr. Prasanna S. Haddela, and my co-supervisor Ms. Thisara Shyamalee for their invaluable guidance, insightful feedback, and unwavering support throughout this research journey. Their expertise and mentorship have played a pivotal role in shaping the direction and quality of this study. I am also sincerely thankful to the academic and technical staff of the Department of Information Technology at the Sri Lanka Institute of Information Technology for providing the necessary resources and an inspiring academic environment to carry out this research. My heartfelt appreciation goes to my fellow students and friends who provided motivation, technical insights, and occasional debugging support during the development of the Unified Bias Metric. Their camaraderie and collaboration made the process both intellectually enriching and personally enjoyable. Special thanks go to my family, whose constant encouragement, patience, and emotional support gave me the strength to persevere through challenging times. Finally, I would like to acknowledge the open-source communities and developers behind the tools and models such as YOLOv8, DepthAnything v2, CLIP, SHAP, and COCO dataset, without which this research would not have been possible.

TABLE OF CONTENT

1. INTRODUCTION	1
1.1. Background & Literature Review	1
1.1.1. Rise of AI in Vision Systems	1
1.1.2. Bias in Visual AI.....	2
1.1.3. What is Contextual Gender Bias?	3
1.1.4. Limitations of Current Tools.....	5
1.1.5. The Need for a Unified Metric.....	6
1.1.6. Literature Review	7
1.2. Research Gap.....	12
1.3. Research Problem.....	13
1.4. Research Objectives	14
2. METHODOLOGY.....	16
2.1. Overview of the Research Framework	16
2.2. Dataset Preparation.....	19
2.2.1. Dataset Selection	19
2.2.2. Object Category Filtering.....	19
2.2.3. Metadata Structuring	20
2.2.4. Final Dataset Composition.....	21
2.3. Object Detection and Segmentation	22
2.3.1. Model Selection & Tools	22
2.3.2. Detection and Segmentation Process	23
2.3.3. Identification of Prominent Person	24

2.3.4.	Output & Storage Format.....	24
2.3.5.	Limitations of Segmentation Models	25
2.4.	Depth Map Generation	26
2.4.1.	Importance of Depth in Contextual Bias Detection	26
2.4.2.	Depth Estimation Using DepthAnything v2	26
2.4.3.	Depth Feature Extraction	27
2.4.4.	3D Distance Calculation	28
2.4.5.	Integration with the UBM Pipeline	28
2.5.	Scene-Level Feature Extraction	29
2.5.1.	Scene Embedding Using CLIP	29
2.5.2.	Construction of Gendered Scene Embedding References	30
2.5.3.	Scene Similarity Bias (SSB) Calculation	30
2.6.	Unified Bias Metric (UBM) Formulation	31
2.6.1.	Object Influence Score (OIS).....	31
2.6.2.	Scene Similarity Bias (SSB)	32
2.6.3.	Final UBM Score	32
2.7.	Comparative Approaches for UBM Computation.....	33
2.7.1.	Objective of Multi-Approach Design.....	34
2.7.2.	Summary of UBM Computational Approaches	34
2.7.3.	Weight Assignment Techniques	35
2.7.4.	Evaluation and Visualization Strategy	35
2.8.	Testing and Implementation	37
2.8.1.	Dataset Structure and Composition.....	38
2.8.2.	Modular Pipeline Design.....	38

2.8.3.	Execution Environment and Output Artifacts.....	39
2.8.4.	Visual Evaluation Strategy.....	39
2.9.	Limitations and Assumptions	41
2.9.1.	Manual Binary Gender Labeling.....	41
2.9.2.	Bias in Pre-Trained Models	41
2.9.3.	Dataset Imbalance and Representation	41
2.9.4.	Fixed Scene and Object Semantics	42
2.9.5.	Threshold-Based Categorization.....	42
2.10.	Commercialization Aspects of the Metric	42
3.	RESULTS & DISCUSSION.....	44
3.1.	Results	44
3.1.1.	Experimental Setup Recap	44
3.1.2.	Bias Score Distributions Across Approaches	46
3.1.3.	Dataset-Wise Behavior Across Bias Types.....	49
3.1.4.	Feature Weighting and Interpretability	53
3.1.5.	Statistical Significance Validation	54
3.2.	Research Findings	55
3.2.1.	Validation-Based Comparison of Approaches.....	55
3.3.	Discussion	61
3.3.1.	Cross-Approach Stability	61
3.3.2.	Limitations of Detection	62
3.3.3.	Interpretation of Bias Trends	64
3.3.4.	Selection of the Final Unified Bias Metric	66
4.	CONCLUSION.....	68

5. REFERENCES.....	70
6. APPENDIX-A.....	73
7. APPENDIX-B.....	74
7.1. Human Survey	74
7.2. Plagiarism Report	75
7.3. Pipeline Structure	75

LIST OF TABLES

Table 1:Sample YOLO + SAM Output Metadata23

Table 2: Summary of UBM Weighting Strategies and Models34

Table 3: Description of Evaluation Datasets Used in the UBM Pipeline38

LIST OF FIGURES

Figure 1.1: Timeline of key milestones in deep learning-based computer vision models, from AlexNet (2012) to Detnas and CLIP, highlighting breakthroughs in image classification, object detection, and multimodal learning	2
Figure 1.2: Example of contextual gender bias: three images of women incorrectly predicted as male. Despite correct labels, the model's predictions appear to be influenced by background context and appearance.	3
Figure 1.3: Misclassification due to contextual bias: A male subject is predicted as female, potentially influenced by background elements and behavior such as skincare or the presence of a mirror. This highlights how latent contextual cues may override subject-centric features in AI classification systems.	4
Figure 2.1: UBM pipeline	18
Figure 2.2: Example JSON Metadata Format.	21
Figure 2.3: Bar Chart of Object Occurrences by Gender	22
Figure 2.4: YOLOv8 Detection Output	25
Figure 2.5: SAM segmentation Image	25
Figure 2.6: Original Image vs. Generated Depth Map	27
Figure 2.7: Overlay of Object Masks with Depth Values	28
Figure 2.8: Object-wise Bias Scores computed using Approach 3. Bias categories are color-coded and sorted by object class.	36
Figure 2.9: KDE Plot of Bias Score Distribution for Approach 3 (SHAP + PCA Fusion). The distribution illustrates clear clustering of male-biased, female-biased, and neutral object classes.	36
Figure 2.10: SHAP feature importance bar plot indicating contributions of individual features toward bias score prediction	37
Figure 2.11: Violin plots visualizing bias score distributions for all nine computational approaches across the eight dataset conditions. The plots reveal clear separation patterns, distribution skewness, and variability in model sensitivity depending on dataset bias (balanced, skewed, or extreme)	40

Figure 3.1: Boxplot of Bias Score Distributions Across Approaches (Original Dataset)	47
Figure 3.2: KDE Distributions of Bias Scores Across Datasets and Approaches	48
Figure 3.3: Boxplot Distribution Across Approaches	50
Figure 3.4: Mean Bias Score Analysis	51
Figure 3.5: Variance Analysis	51
Figure 3.6: Bias Categorization Trends	52
Figure 3.7: Feature Weighting Across SHAP (Approach 2), PCA (Approach 3), and Combined Weights (Approach 9), showing relative importance assigned to 3D Distance, Normalized Depth, and Relative Size in contextual bias computation.	54
Figure 3.8: Agreement between predicted object-level bias and full-image gender misclassification trends.	56
Figure 3.9: Proportion of unique objects per dataset for which each approach computed a final bias score. Higher values indicate full object-wise coverage, regardless of the final bias category (male, female, or neutral).	57
Figure 3.10: Distribution of bias categories (male-biased, female-biased, neutral) across all datasets.	58
Figure 3.11a : Pearson correlation heatmap of bias scores across all approaches.	58
Figure 3.12: Dendrogram showing hierarchical clustering of bias detection approaches based on correlation similarity.	59
Figure 3.13: Line plot of bias scores for the top 10 most diverging objects across approaches.	60
Figure 3.14: Human perception of gender bias across object classes.	61

LIST OF ABBRIVIATIONS

Abbreviation	Description
AI	Artificial Intelligence
UBM	Unified Bias Metric
OIS	Object Influence Score
SSB	Scene Similarity Bias
CLIP	Contrastive Language–Image Pretraining
YOLO	You Only Look Once
SAM	Segment Anything Model
PCA	Principal Component Analysis
SHAP	SHapley Additive exPlanations
SMOTE	Synthetic Minority Over-sampling Technique
XGBoost	eXtreme Gradient Boosting
CSV	Comma-Separated Values
GPU	Graphics Processing Unit
RGB	Red Green Blue
KDE	Kernel Density Estimation

1. INTRODUCTION

1.1. Background & Literature Review

1.1.1. Rise of AI in Vision Systems

The field of computer vision has undergone a revolutionary transformation with the advent of deep learning, particularly convolutional neural networks (CNNs), which have enabled machines to interpret and analyze visual data with human-like precision. This shift was catalyzed by breakthroughs such as AlexNet, which dramatically improved performance in the ImageNet competition and sparked the deep learning revolution in vision tasks [1]. Subsequent models like YOLO (You Only Look Once) introduced real-time object detection capabilities [2], while multimodal models such as CLIP integrated vision and language understanding using massive pre-training on image-text pairs [3].

These advances now power a wide array of real-world applications spanning facial recognition in security systems, diagnostic imaging in healthcare, autonomous navigation in self-driving vehicles, and content moderation on social media platforms. As AI systems are increasingly deployed in critical decision-making processes, the importance of ensuring their reliability, fairness, and ethical behavior has become a paramount concern. These vision models are not merely tools for automation, they influence hiring decisions, medical diagnoses, policing strategies, and media visibility. Any systemic flaw or bias in their predictions, therefore, carries the risk of amplifying existing social inequalities [4]. This has led researchers, developers, and policymakers to critically examine not only the performance metrics of AI models but also their fairness across different demographic and contextual dimensions. As visual AI continues to evolve, its integration into society demands rigorous scrutiny not just for what these systems can do, but how equitably and justly they do it.

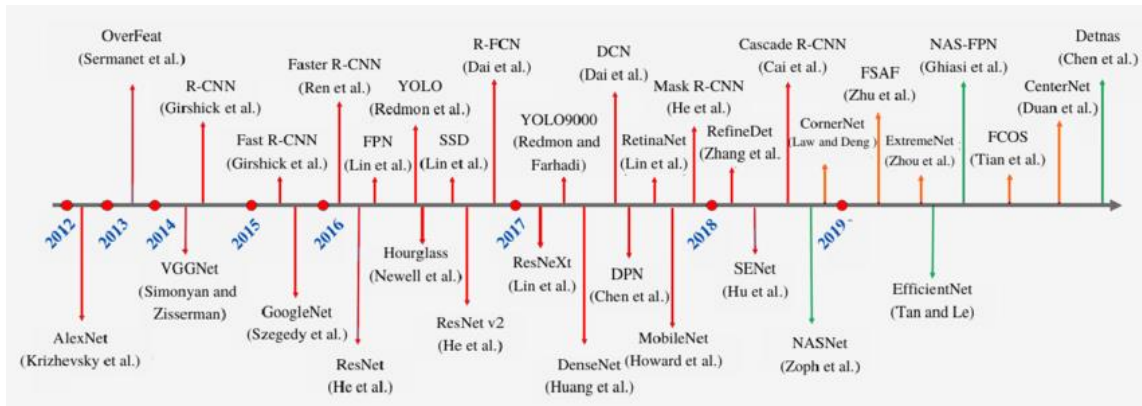


Figure 1.1: Timeline of key milestones in deep learning-based computer vision models, from AlexNet (2012) to Detnas and CLIP, highlighting breakthroughs in image classification, object detection, and multimodal learning.

1.1.2. Bias in Visual AI

While artificial intelligence has shown remarkable capability in automating complex visual tasks, it has also revealed critical shortcomings in fairness, particularly when deployed in socially sensitive contexts. Bias in AI can manifest across multiple levels: demographic bias (where specific groups are underrepresented or misclassified), label bias (inaccuracies due to flawed annotations), representational bias (limited portrayal of certain groups), and most insidiously, contextual bias where background elements of an image subtly affect predictions. In the domain of gender classification, such biases can have tangible consequences, including miss-gendering in facial recognition, skewed recommendations in hiring platforms, and misinterpretations in healthcare diagnostics.

Gender bias in visual models is especially concerning because it often intersects with other dimensions of identity like race, age, and physical appearance. Studies have shown that commercial face analysis tools exhibit the highest error rates for dark-skinned women up to 34.7% in some cases compared to less than 1% for light-skinned men [4]. These disparities raise concerns about the structural inequities being replicated by machine learning models trained on large-scale datasets that may not be demographically representative or contextually balanced.

Moreover, the root causes of such bias often remain obscured by the black-box nature of deep learning models. Visual classifiers frequently rely not just on the person's features,

but also on surrounding visual cues such as objects, scenes, and lighting which may have been incidentally correlated with gender during training. This complicates the mitigation process: balancing the dataset is often not enough, as the model may have internalized latent associations between gender and contextual patterns, such as associating kitchen settings with women or sports fields with men.



Figure 1.2: Example of contextual gender bias: three images of women incorrectly predicted as male. Despite correct labels, the model's predictions appear to be influenced by background context and appearance.

This calls for a shift in focus from solely evaluating demographic parity to quantifying how non-subject elements influence AI decisions a domain still under-explored in mainstream fairness audits.

1.1.3. What is Contextual Gender Bias?

Contextual gender bias refers to the unintended influence of visual context such as objects, spatial layout, background environment, and scene semantics on AI-driven gender classification. Unlike demographic or representational bias, which stem from who is represented in the data, contextual bias originates from how they are visually represented. In this form of bias, a classifier may predict gender not just based on a person's appearance but also due to the surrounding elements in an image that act as latent cues. For instance, a woman pictured in a sports field or holding a skateboard may be classified as male because the model has learned to associate such contexts with masculinity.

This is particularly alarming in cases where datasets are balanced across genders, yet bias persists due to the learned correlations between gender labels and contextual elements. This kind of bias arises from the model’s internal feature representations what it “pays attention to” during prediction. When the visual context includes stereotypically gendered elements (e.g., kitchen utensils, handbags, or sports equipment), the classifier may rely on these instead of facial or bodily features.

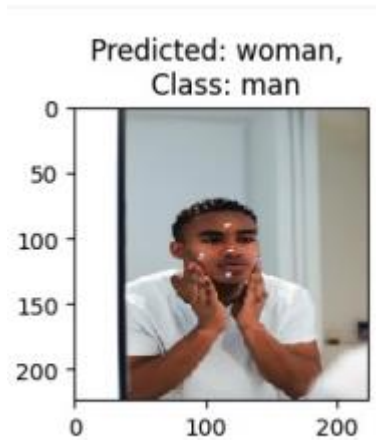


Figure 1.3: Misclassification due to contextual bias: A male subject is predicted as female, potentially influenced by background elements and behavior such as skincare or the presence of a mirror. This highlights how latent contextual cues may override subject- centric features in AI classification systems.

Research by Sabir and Padró [5] showed that objects like lipstick or kitchenware skew gender predictions depending on their relative size and position in the image. Similarly, Meister et al. [6] introduced the notion of “gender artifacts” background features like lighting, scene layout, or co-occurring objects that inadvertently encode gender signals. These artifacts are prevalent in datasets such as COCO and OpenImages, making them difficult to eliminate without losing vital contextual data. As AI becomes increasingly integrated into decision-making systems, failing to account for such context-driven bias poses significant ethical and practical risks.

Contextual gender bias not only challenges the notion of fairness but also complicates existing efforts in dataset balancing and model debiasing. It calls for advanced diagnostic

tools that go beyond simple label statistics and examine the interplay between subject and scene setting the foundation for metrics like the Unified Bias Metric (UBM).

1.1.4. Limitations of Current Tools

Despite the growing focus on fairness in AI, most existing tools for bias detection and mitigation fall short when applied to visual data especially in contexts where object-scene relationships subtly influence predictions. Traditional fairness frameworks tend to emphasize dataset-level co-occurrence statistics, demographic parity, or subject-centric classification accuracy. However, they often overlook the spatial and semantic context that can significantly sway outcomes in computer vision systems.

REVISE is a prominent dataset auditing tool that highlights object-gender co-occurrence patterns to flag representational imbalances. While effective in identifying demographic over- or underrepresentation in datasets like COCO, REVISE lacks spatial sensitivity. For instance, it cannot distinguish between a handbag in the foreground and one in the background despite the fact that object prominence can meaningfully alter gender predictions [7].

AIF360 (IBM) and Fairlearn (Microsoft) offer fairness metrics like statistical parity and disparate impact, but are primarily designed for tabular data and structured inputs. These tools do not analyze visual context, and are therefore ill-suited to diagnose bias arising from background cues, object salience, or spatial proximity in image-based classification tasks [8].

Gender Shades, a pioneering study that revealed intersectional accuracy disparities in facial analysis systems, focuses exclusively on the facial region specifically gender and skin tone accuracy. It does not explore how the surrounding environment or co-occurring objects might influence model outputs, leaving contextual bias unaddressed [4].

Grad-CAM and Score-CAM offer pixel-level interpretability by producing saliency heatmaps, showing which image regions contributed to a model’s prediction. While valuable for qualitative insights, these tools are not designed for quantifying bias across

large datasets. They lack a systematic scoring mechanism for auditing contextual influence [9] [10].

This fragmented landscape highlights the need for a unified, scalable, and explainable metric that can bridge the gap between qualitative heatmaps and quantitative bias assessment. The proposed Unified Bias Metric (UBM) aims to fill this gap integrating object prominence, spatial distance, depth, and scene semantics into a single, interpretable framework for measuring contextual gender bias in image datasets.

Tool	Data Type	Context-Aware	Visual Input	Quantifiable Score
REVISE	Image metadata	No	No	Yes
AIF360	Tabular	No	No	Yes
Gender Shades	Facial images	No	Partial	Yes (but limited)
Grad-CAM	Images	Partial	Yes	No (qualitative only)
UBM (Proposed)	Images to Tabular data	Yes	Yes	Yes

1.1.5. The Need for a Unified Metric

As the demand for fairness in AI intensifies, researchers have proposed a variety of tools and frameworks to detect and mitigate bias in machine learning models. However, a critical shortcoming remains in the domain of visual AI the lack of a unified, quantitative metric that captures contextual bias. Most existing tools have a narrow focus: they either assess dataset-level co-occurrence statistics, such as REVISE [7], or measure facial accuracy disparities, like Gender Shades [4]. While these frameworks are important milestones, they are not designed to analyze how spatial object arrangement or scene semantics affect predictions in real-world visual settings.

Contextual gender bias is inherently multidimensional rooted in object co-occurrence, spatial proximity, depth perception, and scene semantics. For instance, Sabir and Padró

[5] demonstrated that objects such as lipstick or kitchenware influence gender predictions depending on their position and size relative to the subject, while Meister et al. [6] revealed that seemingly neutral background features can function as “gender artifacts” embedded in the scene itself. However, current fairness tools do not combine these diverse influences into a cohesive scoring mechanism.

Moreover, the dominant interpretability tools like Grad-CAM and Score-CAM [9] [10] offer only qualitative visualizations through heatmaps. These methods show where a model is “looking” during prediction but fail to provide numerical bias scores that can be compared across datasets or used for statistical audits. Similarly, widely used fairness libraries like Fairlearn and AIF360 [8] focus on structured or tabular data and lack the semantic understanding needed for image-based contextual analysis.

This gap necessitates a novel metric that is vision-specific, interpretable, and scalable. The Unified Bias Metric (UBM) proposed in this research aims to address these limitations by combining two complementary components: Object Influence Score (OIS), which captures spatial, depth, and size-based object bias, and the Scene Similarity Bias (SSB), which measures the semantic influence of scenes using embeddings from models like CLIP and Places365. Unlike prior tools, UBM leverages SHAP for explainability, supports dataset-wide comparisons, and allows modular extensions for future fairness auditing. By bridging the gap between object-level salience, scene semantics, and visual interpretability, UBM offers the quantitative diagnostic framework tailored specifically to contextual bias in visual gender classification.

1.1.6. Literature Review

The landscape of research on algorithmic bias in visual AI systems has grown significantly, with numerous studies highlighting disparities in how models treat individuals based on race, gender, and other sensitive attributes. While demographic and representational biases have been extensively audited using statistical parity or fairness metrics, a more elusive form contextual bias has recently emerged as a critical concern. Contextual bias refers to the phenomenon where background elements in an image, such

as objects, spatial configurations, or overall scene semantics, influence model predictions in unintended ways. This form of bias is especially insidious because it is latent, difficult to detect, and often persists even in datasets that are balanced demographically.

Researchers have increasingly pointed out that visual classifiers may misclassify gender not due to inherent subject features, but because of correlated environmental cues a woman holding a skateboard might be labeled male due to the stereotypical context of sports, for example. Despite these insights, most bias detection tools and explainability frameworks fall short in systematically quantifying this kind of bias. This literature review explores key findings from past research on object-driven and scene-driven bias, evaluates existing interpretability tools, and identifies gaps that justify the development of a Unified Bias Metric (UBM). UBM is specifically designed to measure contextual gender bias by integrating object-level spatial features and scene-level semantic embeddings, thereby addressing limitations in current fairness toolkits and interpretability models.

1.1.6.1. Object Bias & Spatial Features

Bias in visual AI systems often emerges not just from subject characteristics but from the surrounding objects and their spatial arrangements. These contextual cues can significantly influence gender predictions, especially in scenarios where the model implicitly learns associations between gender and certain visual elements. Sabir and Padró [5] conducted a detailed analysis of how objects such as lipstick and kitchenware typically associated with female gender roles bias the output of captioning systems and image classifiers. Their work introduced a semantic-based gender score to quantify the degree of object-associated gender bias and showed that the size and prominence of these objects within the image markedly skew model predictions. The implications of their findings support the importance of modeling object salience and spatial prominence when evaluating fairness in AI.

Expanding on this concept, Bhargava and Forsyth [11] demonstrated that image captioning systems tend to produce gendered descriptions based on contextual objects like skateboards or ovens. Their work revealed that even when datasets are balanced in terms

of gender, the surrounding objects still heavily influence the language generated by the model. To mitigate such effects, they proposed a modular decoupling framework that first generates a gender-neutral caption and then independently classifies gender, thereby isolating subject features from context. This methodological separation underscores the necessity for explainable models that account for contextual influences at the object level.

Together, these studies highlight that traditional bias detection mechanisms those focusing solely on subject attributes are insufficient for diagnosing deeper contextual dependencies. The concept of Object Influence Score (OIS), proposed in this research, builds upon this foundation by incorporating object size, spatial distance, and depth to quantify how much influence a co-occurring object exerts on gender classification. This lays the groundwork for a more interpretable approach to measuring contextual bias in visual AI systems.

1.1.6.2 Scene Semantics & Gender Artifacts

While much of the bias research in computer vision has centered on the subject or co-occurring objects, emerging work has revealed the powerful role of scene semantics in shaping model predictions. Scene semantics refer to the holistic understanding of a visual environment including its background setting, spatial layout, lighting, and symbolic context which AI models increasingly use, sometimes erroneously, as cues in classification tasks. Meister et al. [6] introduced the notion of gender artifacts, subtle background elements such as lighting, scene composition, and object placement that carry implicit gender associations. Their study showed that even after controlling for subject-level demographic and labeling biases, vision models continued to rely on these background cues, often resulting in systematic gender misclassification.

For example, in the COCO and OpenImages datasets, scenes depicting sports fields, garages, or urban streets were more likely to lead models toward male predictions, whereas kitchen interiors, soft-lit bedrooms, or domestic environments skewed predictions toward female. These findings underscore how contextual bias can emerge not only from foreground objects but from the broader visual backdrop in which the subject is embedded.

To operationalize scene-level analysis, modern semantic embedding models like CLIP [3] and Places365 [12] have become invaluable. CLIP transforms images into high-dimensional vectors aligned with natural language, capturing semantic nuances about a scene’s content. Meanwhile, Places365 classifies images into one of 365 scene categories (e.g., classroom, living room, sports arena), offering a structured lens into environmental semantics. These tools enable the computation of the Scene Similarity Bias (SSB), a key component of the proposed Unified Bias Metric (UBM), by measuring how closely an image’s scene matches environments typically associated with specific genders. This embedding-based analysis represents a novel direction in fairness research, moving beyond object co-occurrence to quantify scene-contextual influences in large-scale image datasets.

1.1.6.3. Explainability vs. Quantification

As concerns around fairness in AI systems grow, so too does the demand for interpretable models. In computer vision, tools such as Grad-CAM [9] and Score-CAM [10] have emerged as popular methods for visualizing which regions of an image contribute to a model’s prediction. These heat-map based methods highlight areas of focus within a neural network, offering insights into whether a model is attending to relevant parts of the image (e.g., the face of a person) or being influenced by irrelevant or biased context (e.g., background objects like kitchen appliances). While these tools provide powerful qualitative insights, they are inherently limited in their ability to quantify bias across large datasets or compare bias scores between models.

This limitation presents a major challenge in auditing contextual bias at scale. Grad-CAM and Score-CAM produce instance-specific saliency maps, which are useful for inspection but not for statistical validation or cross-sample analysis. They also do not provide numerical outputs that can be tracked or aggregated, which restricts their utility in dataset-wide fairness audits or in comparative benchmarking of models. As a result, these tools fall short in settings where systematic, explainable, and scalable quantification is essential particularly in assessing contextual bias in gender classification.

To address this gap, the Unified Bias Metric (UBM) integrates SHAP (SHapley Additive exPlanations) as a core component of its scoring pipeline. Unlike Grad-CAM or Score-CAM, SHAP provides feature-level attributions in a mathematically grounded, model-agnostic manner, allowing for direct comparisons across instances, models, and datasets. By assigning each contextual feature (such as object proximity, size, depth, or scene embedding similarity) a contribution value to the model’s prediction, SHAP enables explainability that is both interpretable and quantifiable. This makes it ideally suited for fairness diagnostics where transparency, accountability, and auditability are critical.

UBM’s use of SHAP positions it as a novel framework that bridges the gap between traditional visual explanation tools and rigorous, bias-aware metrics. It enhances interpretability not just through visualization, but through structured, numerical evaluation marking a crucial step forward in explainable fairness auditing for visual AI systems.

1.1.6.4. Summary of Gaps and Justification for UBM.

Despite the rapid evolution of fairness-focused tools in visual AI, a solution for quantifying contextual gender bias remains elusive. Existing methods typically target either demographic parity (as seen in Fairlearn and AIF360 [8]) or focus on limited co-occurrence analysis (e.g., REVISE [7]), without accounting for spatial relationships or scene semantics. Even more targeted tools like Gender Shades [4] concentrate on subject appearance and fail to consider how context may shape model decisions. Interpretability methods such as Grad-CAM [9] and Score-CAM [10] are helpful in localizing attention, yet lack the infrastructure for scalable, dataset-level quantification of bias especially when that bias is embedded in objects or scenes rather than the subject itself.

Recent works, including those by Sabir and Padró [5], Meister et al. [6], and Bhargava & Forsyth [11], show that both object proximity and scene semantics play critical roles in shaping gender predictions. However, no existing framework synthesizes these elements into a single interpretable metric. Furthermore, tools like CLIP [3] and Places365 [12] offer embeddings for semantic comparison but are rarely integrated into fairness pipelines for quantifying scene influence.

The Unified Bias Metric (UBM) is proposed to address this multidimensional challenge. It is the framework to systematically incorporate object-level features (size, spatial distance, depth), scene-level semantics (via CLIP and Places365 embeddings), and explainability (through SHAP) into a quantitative, auditable metric. UBM enables object-wise and image-wise diagnosis of contextual bias and supports statistical analysis across datasets bridging a critical gap in current fairness auditing. This innovation sets the foundation for the methodology outlined in the following chapter.

UBM further enables per-object bias scoring by aggregating contextual features across each object class, allowing researchers to quantify how strongly specific objects (e.g., handbags, sports balls) influence gender classification outcomes.

1.2. Research Gap

Despite the rapid advancement in fairness-aware machine learning, there remains a critical and underexplored challenge in the domain of computer vision: the lack of tools capable of diagnosing and quantifying contextual gender bias. Existing fairness toolkits like REVISE [7], Fairlearn [8], and AI Fairness 360 [8] primarily focus on demographic imbalances, label co-occurrences, or overall accuracy disparities, but they fail to assess the subtle and often latent biases arising from visual context. These tools are often designed for structured, tabular data or subject-centric facial analysis and overlook the influence of background elements such as spatial object positioning, scene semantics, and object depth.

While studies like those by Sabir and Padró [5] and Meister et al. [6] have highlighted how object size, depth, and surrounding environments contribute to gender misclassification, their insights remain largely theoretical or qualitative. There is still no unified metric that systematically integrates object-level salience with scene-level semantics to produce numerical bias scores for large-scale auditing. Moreover, popular interpretability tools like Grad-CAM and Score-CAM [9][10] offer only image-specific heatmaps and lack generalizability for dataset-wide fairness evaluation.

Furthermore, the use of semantic embedding models like CLIP [3] and Places365 [12] for fairness diagnostics has not been formally integrated into bias metrics. These models possess strong capabilities for capturing visual-language alignment and environmental context but are rarely leveraged to measure scene-gender associations in a quantifiable way.

This gap highlights the urgent need for a new diagnostic framework that is image-specific, explainable, statistically grounded, and capable of evaluating both object influence and scene context. Without such a framework, current fairness assessments remain incomplete, and AI systems continue to risk perpetuating bias through latent visual correlations that are neither labeled nor obvious during training.

1.3. Research Problem

Despite the growing recognition of fairness challenges in artificial intelligence (AI), most existing tools for bias detection in visual datasets remain limited to label-level statistics (e.g., co-occurrence frequencies) or demographic group disparities in subject-focused models. These approaches overlook a critical dimension of bias contextual bias where the presence of surrounding objects, their spatial relationships, and the broader scene environment inadvertently influence gender classification outcomes. Although prior studies have acknowledged the role of object-gender associations and scene-induced misclassifications, there is still no unified and interpretable metric capable of quantifying these contextual influences in a systematic and scalable manner.

This research seeks to address the problem of how to detect and measure contextual gender bias in image datasets, particularly in scenarios where traditional demographic analyses and co-occurrence-based metrics prove insufficient. The core issue lies in the current lack of tools that can integrate object-level properties such as size, depth, and spatial distance with scene-level semantics derived from embedding models like CLIP or Places365 into a single, interpretable framework. Furthermore, most explainability tools (e.g., Grad-CAM, Score-CAM) offer only qualitative insights through heatmaps, falling short of

delivering structured, comparative scores that could facilitate large-scale bias audits or benchmarking efforts.

Therefore, the central problem addressed by this study is:

“How can we develop a unified metric that explains and quantifies the influence of visual context comprising object prominence, spatial layout, depth, and scene semantics on gender classification in image datasets?”

This problem sits at the intersection of fairness, explainability, and computer vision. Addressing it would lay the groundwork for a robust, scalable, and interpretable fairness auditing framework capable of diagnosing contextual gender bias in image-based AI systems.

1.4. Research Objectives

To address the challenges outlined in the research problem, this study sets forth a series of targeted objectives aimed at designing, implementing, and validating a unified metric for contextual gender bias in image datasets. The overarching goal is to develop a quantifiable and interpretable framework that integrates both object-level and scene-level visual context into bias evaluation.

Specific Objectives:

- To define and formalize contextual gender bias in the domain of image classification, particularly highlighting its distinction from demographic, label, and representational bias.
- To develop a novel Unified Bias Metric (UBM) that quantifies the influence of visual context on gender classification outcomes, by:
 - Designing the Object Influence Score (OIS) using features such as object size, spatial distance from the subject, and depth.
 - Designing the Scene Similarity Bias (SSB) using semantic scene embeddings from pre-trained models like CLIP and Places365.

- To implement and compare multiple computational approaches (e.g., PCA, SHAP, XGBoost, Ridge Regression) for UBM calculation, ensuring interpretability and scalability.
- To construct a curated, gender-labeled image dataset by filtering and annotating images from the COCO dataset based on object presence and scene types to isolate contextual features.
- To evaluate the effectiveness of UBM across multiple biased and balanced datasets, and validate its ability to uncover contextual biases undetectable by traditional fairness metrics.
- To visualize and explain the contextual bias scores using interpretable methods such as SHAP values, embedding similarity plots, and comparative violin/box plots by gender.

2. METHODOLOGY

2.1. Overview of the Research Framework

This study proposes a structured, multi-stage methodology to detect and quantify contextual gender bias in visual datasets through a novel metric called the Unified Bias Metric (UBM). The framework is designed to measure how both object-level features (e.g., size, spatial positioning, depth) and scene-level semantics (e.g., background context) influence AI-driven gender classification outcomes. In contrast to traditional fairness evaluations that rely on co-occurrence or demographic group statistics, this approach captures contextual dependencies by analyzing how individuals and objects are visually arranged within each scene.

The UBM pipeline consists of four main stages: data preparation, feature extraction, bias scoring, and interpretability analysis. Its design prioritizes scalability, reproducibility, and extensibility. At its core, the metric integrates two principal components:

- **Object Influence Score (OIS):** Quantifies the influence of object-level spatial features including relative size, 3D proximity, and depth difference on gender classification.
- **Scene Similarity Bias (SSB):** Measures the semantic alignment of image scenes with male- or female-dominant contexts using pre-trained scene embeddings.

The final contextual bias score is calculated as a weighted combination of OIS and SSB:

$$UBM = \alpha * OIS + \beta * SSB$$

where the weights (α , β) are obtained through dimensionality reduction techniques (e.g., PCA), explainability models (e.g., SHAP), or predictive regression models (e.g., Ridge).

Framework Pipeline Overview

The complete pipeline consists of the following steps:

1. Dataset Curation and Gender Labeling

A filtered subset of the Microsoft COCO dataset [13] is selected, focusing on images where a person co-occurs with objects commonly associated with gender stereotypes. Binary gender labels are manually annotated for the most prominent person in each image.

2. Object Detection and Segmentation

Persons and objects are detected using YOLOv8 (Ultralytics, 2023) [14]. For improved segmentation precision, masks are refined using the Segment Anything Model (SAM) [15]. Bounding boxes, masks, and object labels are stored in JSON metadata for downstream analysis.

3. Depth Estimation and 3D Contextual Analysis

Depth maps are generated using the DepthAnything v2 model [16], enabling the computation of normalized object-person distances in both 2D and 3D space. These features feed into the OIS component.

4. Scene Embedding and Contextual Bias Evaluation

Semantic embeddings are extracted using CLIP (OpenAI, 2021) [3] and optionally supported with Places365 [12] scene classifications. Cosine similarity is computed between image scene embeddings and gender-reference scene vectors to produce the SSB.

5. Feature Normalization and Engineering

All extracted features relative size, 3D distance, depth, and scene similarity are normalized to ensure comparability across images of varying scale and resolution.

6. UBM Score Computation and Weight Optimization

The UBM score is computed using multiple weighting strategies across nine experimental approaches:

- PCA-based variance weighting
- SHAP-based feature attribution
- XGBoost and Ridge regression modeling

- Optuna-optimized model tuning

These variations allow for evaluating robustness and sensitivity under different learning assumptions.

7. Interpretability and Visualization

Visual outputs include:

- SHAP plots for feature influence
- Bar and violin plots for object-level bias scores
- Scene embedding similarity maps

All results and visualizations are generated using Python (matplotlib, seaborn) and stored as CSV files for reproducibility.

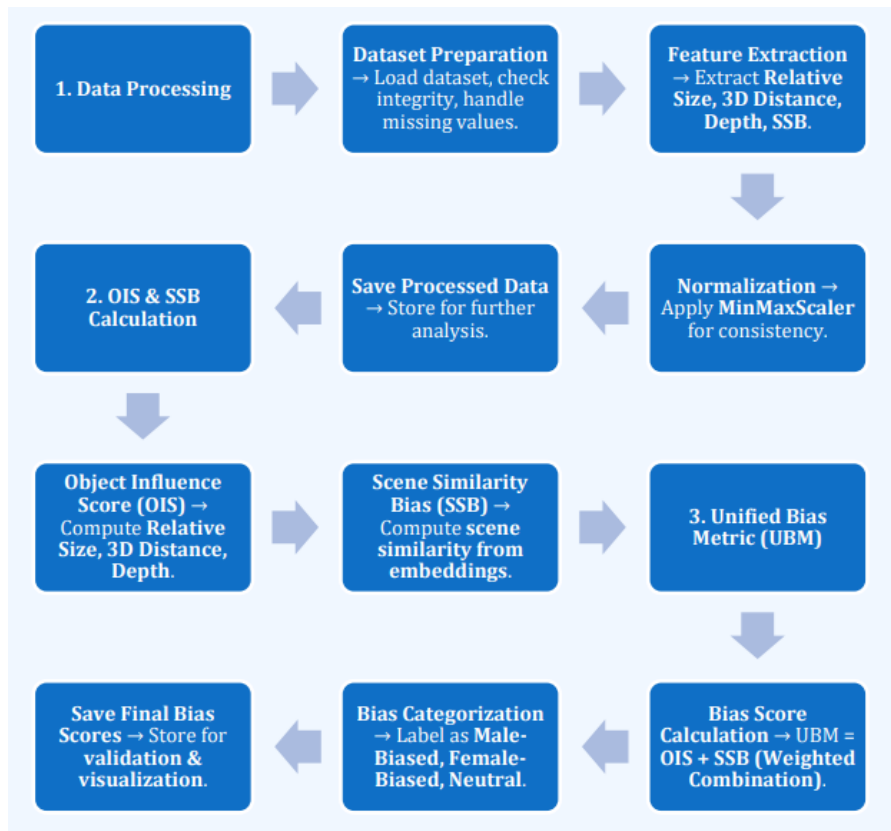


Figure 2.1: UBM pipeline

This methodology offers a modular, extensible, and reproducible pipeline for measuring contextual bias in AI vision systems. By combining explainable machine learning, semantic embeddings, and spatial modeling, UBM provides a tool for auditing gender bias in image datasets across diverse domains such as surveillance, hiring, and content generation.

2.2. Dataset Preparation

The effectiveness of the Unified Bias Metric (UBM) relies heavily on the quality and structure of the dataset used to capture contextual gender bias. To ensure that the metric evaluates images with meaningful object-person interactions and contextual diversity, a curated subset of the Microsoft COCO dataset [13] was selected. The preparation process involved systematic filtering, manual labeling, and metadata structuring, with an emphasis on gender-associated object co-occurrences.

2.2.1. Dataset Selection

The Microsoft COCO dataset was selected due to its comprehensive annotations, diversity of visual contexts, and the availability of object and person instances in a wide range of everyday scenes. For this research, the filtering criterion required that each image must contain at least one person to study how visual context (e.g., objects, background) might influence gender classification. This subset is especially valuable for analyzing contextual biases, as it reflects real-world co-occurrences between people and objects in various environments.

2.2.2. Object Category Filtering

For target potential contextual gender biases, eight object categories were selected based on prior research [4][5][6] and sociocultural stereotypes. These categories were divided as follows:

- Female-associated objects: handbag, hair drier, umbrella, cup
- Male-associated objects: sports ball, baseball bat, bicycle, skateboard

Using the COCO annotations file (instances_val2017.json), a custom Python script filtered out images that featured at least one person and one of the above objects. This step ensured that the final dataset reflected a balance of male- and female-associated contexts likely to trigger bias in AI vision models.

2.2.1. Manual Gender Labeling of Persons

Because the COCO dataset does not include gender annotations, manual labeling was performed. The labeling focused on the most prominent person in each image defined by the largest bounding box area and central positioning.

To improve reliability:

- Two annotators independently assigned binary gender labels (male/female).
- Inconsistencies were resolved through discussion or exclusion.
- Only one gender label was assigned per image to ensure consistent analysis.

These gender labels served as ground truth for evaluating object-level bias in the UBM pipeline.

2.2.3. Metadata Structuring

For each image, a structured JSON file was created to store metadata, which included:

- Image ID and file name
- Bounding boxes and class names of detected objects
- Bounding box of the labeled person
- Manually annotated gender label

Placeholders for computed features such as object size, spatial distance, depth, and scene similarity.

```

{
  "class_name": "person",
  "bbox": [
    172.33811950683594,
    156.70892333984375,
    481.0,
    637.9450073242188
  ],
  "confidence": 0.8721827268600464,
  "mask_path":
"/content/drive/MyDrive/ContextualBiasProject/outputs/all_images/masks/000000000036/mask_person_1.png",
  "mean_depth": 130.99079587429102,
  "normalized_depth": 51.3689395585455,
  "gender": "Female"
},
{
  "class_name": "umbrella",
  "bbox": [
    5.67034912109375,
    56.59185791015625,
    453.52667236328125,
    526.848876953125
  ],
  "confidence": 0.8918861746788025,
  "mask_path":
"/content/drive/MyDrive/ContextualBiasProject/outputs/all_images/masks/000000000036/mask_umbrella_0.png",
  "mean_depth": 118.64993351063829,
  "normalized_depth": 46.52938569044639,
  "relative_size": 1.41785825160928,
  "normalized_distance_xy": 0.000810992856981018,
  "normalized_distance_z": 25.44449020520048,
  "3d distance": 25.444490218124876
}

```

Figure 2.2: Example JSON Metadata Format.

2.2.4. Final Dataset Composition

After filtering and annotation, the final dataset included **852 images**, evenly distributed across the selected object categories. Gender labels were approximately balanced, ensuring meaningful bias detection across diverse visual contexts. This finalized dataset was passed through the UBM pipeline for feature extraction, contextual analysis, and metric computation.

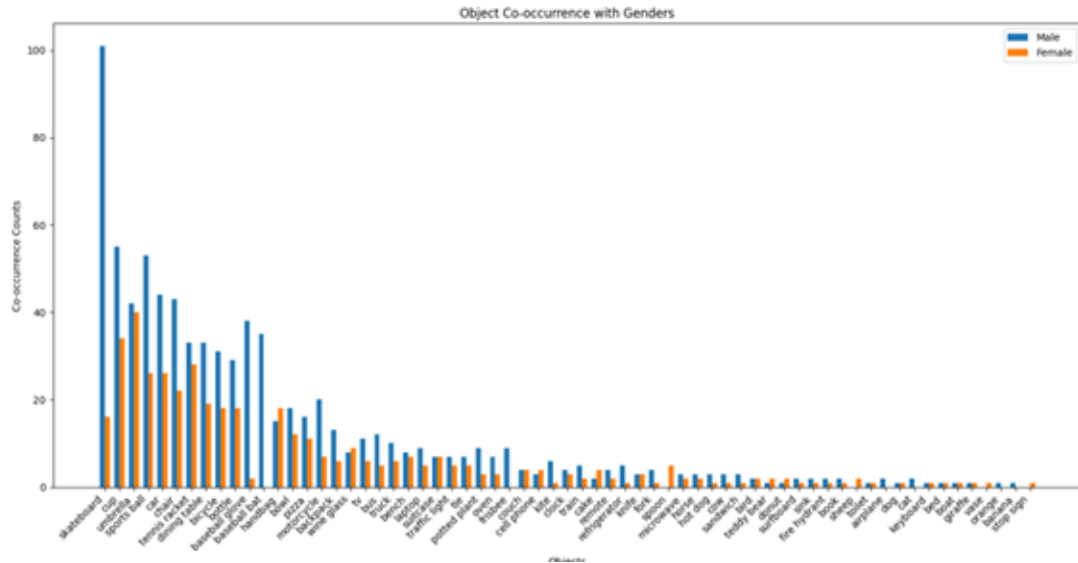


Figure 2.3: Bar Chart of Object Occurrences by Gender

2.3. Object Detection and Segmentation

The accurate detection and segmentation of persons and objects within each image is a foundational step in the Unified Bias Metric (UBM) pipeline. This phase ensures that all visually relevant elements particularly those associated with gender stereotypes are identified, localized, and structured for further analysis such as object-person spatial relationships and feature extraction. High-quality object detection directly influences the reliability of the Object Influence Score (OIS), a critical subcomponent of the UBM.

2.3.1. Model Selection & Tools

To achieve robust and precise detection, two state-of-the-art computer vision models were integrated into the pipeline:

- YOLOv8 (You Only Look Once, Version 8) [14]: A real-time, high-performance object detection model by Ultralytics, used to detect all COCO-labeled objects including persons. YOLOv8 was selected for its lightweight architecture, superior inference speed, and high accuracy across varied object categories.
- Segment Anything Model (SAM) [15]: An advanced segmentation model used optionally to enhance YOLO-generated bounding boxes by producing fine-

grained, pixel-wise masks. This refinement step is especially useful for irregular or overlapping objects (e.g., umbrellas, cups), providing more accurate area estimations necessary for size-based bias computation.

2.3.2. Detection and Segmentation Process

Each image is first passed through YOLOv8 to extract:

- Bounding boxes for all persons and detected objects.
- Class labels based on COCO object taxonomy.
- Confidence scores, filtered using a threshold of ≥ 0.25 .

For segmentation refinement, detected bounding boxes are fed into SAM, which outputs pixel-wise masks for improved spatial delineation. These masks are then linked to the corresponding object metadata.

Metadata for each object includes:

Table 1: Sample YOLO + SAM Output Metadata

Field	Description
Class Label	COCO category (eg : handbag, person)
Bounding Box	Coordinates(x, y, width, height)
Area	Calculated from mask or bounding box
Confidence Score	Detection confidence (range : 0 – 1)
Mask ID	Index for linking to segmentation output

Relative Object Size Calculation

For each detected object, a **relative size score** is computed to reflect its visual prominence. This is calculated as the ratio of the object's bounding box area to the full image area. If segmentation masks from SAM are available, the pixel-wise mask area is used instead for higher accuracy, especially for irregularly shaped or partially occluded objects.

All relative size values are **normalized** across the dataset to maintain consistency and enable comparative analysis across images of varying dimensions. This feature is crucial because larger or foreground objects tend to dominate the visual field and may disproportionately influence gender predictions particularly when such objects are stereotypically gendered (e.g., handbags, sports balls, hair dryers). Relative size is one of the three core components used in computing the **Object Influence Score (OIS)**.

2.3.3. Identification of Prominent Person

To standardize gender annotation and contextual analysis, only the most prominent person per image is retained. Prominence is determined using:

- Largest bounding box area, and
- Proximity to image center (Euclidean distance)

This approach ensures consistency in measuring object influence relative to a single subject per image.

2.3.4. Output & Storage Format

Each processed image produces a dedicated metadata JSON file, storing:

- Image ID and filename
- Bounding boxes and labels of all detected objects
- Bounding box and center-point of the main person
- Optional segmentation mask references (if SAM is enabled)



Figure 2.4: YOLOv8 Detection Output



Figure 2.5: SAM segmentation Image

2.3.5. Limitations of Segmentation Models

While SAM significantly improves spatial granularity, its performance may degrade in cases involving:

- Heavily occluded objects
- Overlapping object masks
- Low-contrast scenes or ambiguous boundaries

Acknowledging these limitations allows for nuanced interpretation of object influence, particularly when comparing between bounding-box-only and refined-mask pipelines.

This module ensures that all gender-relevant visual cues are captured with high fidelity, providing the structural input for downstream spatial, depth, and contextual bias analysis. Together, YOLOv8 and SAM create a hybrid detection system that balances speed, accuracy, and segmentation detail, setting the foundation for explainable and interpretable bias quantification.

2.4. Depth Map Generation

Depth estimation is a crucial component of the Unified Bias Metric (UBM) framework, particularly for computing the Object Influence Score (OIS). While traditional 2D object features like size and position offer insight into visual prominence, incorporating depth enables a more accurate representation of how close or far an object is from the subject (person) in 3D space. This is especially important for evaluating contextual bias, as foreground objects often carry more visual weight in classification systems.

2.4.1. Importance of Depth in Contextual Bias Detection

In both human and machine vision, objects appearing closer are typically more salient and exert a stronger influence on perception. In the context of gender bias, for example, a nearby object such as a handbag or sports ball may significantly skew a model’s gender prediction, while a similar object in the background may have negligible effect.

Integrating depth into the UBM pipeline:

- Enhances contextual proximity detection
- Distinguishes influential cues from background noise
- Increases interpretability and reliability of bias scores

2.4.2. Depth Estimation Using DepthAnything v2

To achieve accurate depth prediction, the study employs DepthAnything v2, a state-of-the-art monocular depth estimation model [16]. Trained on large-scale unlabeled datasets, it produces smooth, high-resolution grayscale depth maps from a single RGB input.

- Brighter regions: Indicate closer objects
- Darker regions: Represent distant background

Each image in the dataset is processed through this model, generating depth maps stored alongside original images for downstream analysis.



Figure 2.6: Original Image vs. Generated Depth Map

2.4.3. Depth Feature Extraction

Once the depth maps are generated, depth-based features are extracted for each detected object using its mask or bounding box. The key features include:

- Mean depth of the object (D_{object})
- Mean depth of the person (D_{person})
- Z-distance (relative depth):

$$Z_{distance} = D_{object} - D_{person}$$

All values are normalized to a $[0, 1]$ range to ensure comparability across lighting conditions and contrast variances.



Figure 2.7: Overlay of Object Masks with Depth Values

2.4.4. 3D Distance Calculation

To compute the spatial relationship between each object and the person in 3D space, the full **Euclidean distance** is calculated using:

$$3D\ Distance = \sqrt{(x_2 - x_1)^2 + (y_2 - y_1)^2 + (z_2 - z_1)^2}$$

Where:

- (x_1, y_1) : Center of the person's bounding box
- (x_2, y_2) : Center of the object's bounding box
- (z_1, z_2) : Mean depth values for the person and object respectively

This 3D distance is a critical component of OIS, helping distinguish between contextually important foreground and less relevant background elements.

2.4.5. Integration with the UBM Pipeline

All depth-related features are added to each image's metadata (in JSON format) and then passed to:

- Object Influence Score (OIS) computation
- UBM score calculation (OIS + SSB)

- Explainability models (e.g., SHAP, PCA)

This integration ensures spatial awareness is incorporated at every level of the pipeline.

By including high-fidelity depth estimation via DepthAnything v2, the UBM framework benefits from improved 3D modeling of object-person interactions. This added dimensionality enhances the robustness, fairness, and interpretability of contextual bias measurement in visual AI systems.

2.5. Scene-Level Feature Extraction

While object-level features offer insight into localized bias (e.g., object size or proximity to the subject), scene-level context captures global semantics such as environment, spatial arrangement, and lighting conditions. These latent cues often ignored by traditional bias audits can significantly influence AI gender classification outcomes. In the Unified Bias Metric (UBM) framework, this broader context is captured through the **Scene Similarity Bias (SSB)**: a score that reflects how semantically “masculine” or “feminine” a scene appears based on pre-trained vision-language associations.

2.5.1. Scene Embedding Using CLIP

To extract high-level scene semantics, each image is passed through CLIP (Contrastive Language–Image Pretraining) [3], a multi-modal model trained on 400M image-text pairs. CLIP encodes each image into a 512-dimensional embedding vector $E_{image} \in \mathbb{R}^{512}$, which captures global visual features—layout, textures, lighting, and compositional structure without requiring explicit object segmentation.

CLIP was selected for:

- Robust generalization across diverse domains
- Pre-training on large-scale paired image-text data
- Compatibility with cosine similarity-based comparison

These embeddings serve as the foundation for scene-gender alignment analysis.

2.5.2. Construction of Gendered Scene Embedding References

To compute SSB, two reference vectors are built by averaging the CLIP embeddings of hand-selected gender-associated scenes:

- E_{male} : Mean embedding of scenes stereotypically associated with males. (e.g., **garage, stadium, gym, workshop**)
- E_{female} : Mean embedding of scenes associated with females. (e.g., **kitchen, dressing room, nursery, living room**)

Formally:

$$E_{male} = \frac{1}{N} \sum_{i=1}^N CLIP(image_{male}, i), E_{female} = \frac{1}{M} \sum_{j=1}^M CLIP(image_{female}, j)$$

These reference vectors provide semantic anchors for gender-context comparison.

2.5.3. Scene Similarity Bias (SSB) Calculation

Each image's embedding is compared against both reference embeddings using cosine similarity:

$$SSB = \cos(E_{image}, E_{male}) - \cos(E_{image}, E_{female})$$

Where:

- $SSB > 0$: Scene is more similar to male-associated environments
- $SSB < 0$: Scene aligns more with female-associated environments
- $SSB \approx 0$: Scene is contextually neutral

The computed SSB is saved in the image's metadata and combined with OIS during final UBM computation:

$$UBM = \alpha \times OIS + \beta \times SSB$$

Where α and β are weights derived via PCA, SHAP, or regression tuning strategies.

Scene-level feature extraction is crucial to capturing latent contextual bias that may not be reflected in objects alone. The Scene Similarity Bias (SSB), computed via CLIP embeddings, enables UBM to quantify global environmental influence. By comparing scene semantics against reference vectors, this method delivers an explainable and reproducible bias score that extends the fairness analysis beyond subject-centric methods.

2.6. Unified Bias Metric (UBM) Formulation

The Unified Bias Metric (UBM) is the central formulation of this research, designed to quantify contextual gender bias in images by integrating both object-level and scene-level features. Unlike traditional metrics that rely solely on demographic co-occurrence or class imbalance, UBM captures how visual context through object salience, spatial dynamics, and semantic background influences gender classification outcomes. The metric is computed per image and can be aggregated across object categories or datasets for large-scale bias audits.

UBM consists of two core components:

- **Object Influence Score (OIS)** — Captures the spatial prominence and salience of gendered objects relative to the person.
- **Scene Similarity Bias (SSB)** — Captures semantic similarity between the scene and gender-associated environments.

2.6.1. Object Influence Score (OIS)

The Object Influence Score (OIS) represents the degree to which a particular object is visually dominant or likely to influence model perception of the subject. It is computed as a weighted combination of three normalized features:

- **Relative Size (S)**: Ratio of object area to image area (or from segmentation mask)
- **3D Distance (D)**: Euclidean distance from the object to the person using spatial + depth coordinates
- **Normalized Depth (Z)**: Mean depth value of the object relative to the subject

The general form of OIS is given as:

$$OIS = \omega_1 \times S + \omega_2 \times D + \omega_3 \times Z$$

Where:

ω_1 , ω_2 , ω_3 are feature weights determined using:

- PCA (variance-based weighting)
- SHAP (model explainability-based importance)
- Ridge Regression or XGBoost (predictive weight learning)

Each object’s OIS score reflects its visual and spatial influence within the image.

2.6.2. Scene Similarity Bias (SSB)

As detailed in Section 2.5, the Scene Similarity Bias (SSB) captures how semantically “masculine” or “feminine” the image’s scene is, based on CLIP embeddings.

SSB is computed using cosine similarity between the scene embedding E_{image} and two reference vectors:

E_{male} and E_{female} .

$$SSB = \cos(E_{image}, E_{male}) - \cos(E_{image}, E_{female})$$

- **Positive SSB:** Scene is more aligned with masculine environments
- **Negative SSB:** Scene is more aligned with feminine environments
- **SSB ≈ 0 :** Scene is neutral or ambiguous

This score allows the metric to incorporate **non-object contextual cues**, which often play an overlooked but critical role in biased predictions.

2.6.3. Final UBM Score

The final Unified Bias Metric (UBM) combines both components as a **weighted linear combination**:

$$UBM = \alpha \times OIS + \beta \times SSB$$

Where:

- α and β are scalar weights assigned to the object and scene components respectively
- These can be:
 - **Fixed values** (e.g., $\alpha = \beta = 0.5$)
 - **Optimized via grid search or hyper-parameter tuning**
 - **Learned through regression models like Ridge or XGBoost**
 - **Derived from SHAP-based feature attribution**

This formulation ensures that both localized object-level signals and holistic scene-level context are accounted for in a single, interpretable metric. The UBM can be computed:

- **Per object–person pair**
- **Per image** (aggregated if multiple objects)
- **Per object class** (e.g., average UBM for “handbag”)
- **Per dataset** (e.g., global bias score)

The Unified Bias Metric (UBM) offers a novel and flexible approach to diagnosing contextual gender bias in visual datasets. By integrating the spatial influence of gender-associated objects and the semantic alignment of the scene environment, UBM provides a more nuanced, data-driven alternative to traditional bias metrics. Its interpretable, modular design supports extensive experimentation and allows for integration with existing fairness auditing workflows.

2.7. Comparative Approaches for UBM Computation

To enhance the robustness, generalizability, and interpretability of the Unified Bias Metric (UBM), this research implements and evaluates nine distinct computational approaches. Each approach combines different weighting strategies ranging from statistical modeling to explainability frameworks to integrate object- and scene-level features into final bias

scores. This analysis offers multiple perspectives on how gender-related contextual bias manifests in visual datasets and enables validation of the stability and fairness of UBM.

2.7.1. Objective of Multi-Approach Design

Since contextual gender bias emerges through both spatial proximity (object-level) and semantic similarity (scene-level), a one-size-fits-all metric may not suffice across datasets or tasks. Therefore, this study adopts a multi-model strategy, where each approach uses a different technique to assign weights to the Object Influence Score (OIS) and Scene Similarity Bias (SSB), including:

- **Unsupervised variance analysis (PCA)**
- **Model-based interpretability (SHAP)**
- **Regularized linear models (Ridge Regression)**
- **Hyperparameter-tuned ensemble methods (XGBoost + Optuna)**

This diversity enables a more nuanced and generalizable interpretation of bias while also improving transparency in how the scores are derived.

2.7.2. Summary of UBM Computational Approaches

Table 2: Summary of UBM Weighting Strategies and Models

Approach	Weighting Strategy	Model Used	Highlights
1	PCA / Random Forest	Regression	Baseline using unsupervised feature variance
2	SHAP Values	Random Forest Classifier	Feature attributions from decision paths
3	SHAP + PCA (Averaged)	RF + PCA	Blends interpretable and unsupervised insights
4	SHAP	XGBoost Classifier	Strong baseline with high explainability

5	SHAP	XGBoost SMOTE	+	Handles gender label imbalance using synthetic oversampling
6	SHAP + PCA + Grid Search	XGBoost SMOTE	+	Optimizes blend of SHAP-PCA weights via exhaustive search
7	SHAP	XGBoost Optuna	+	Hyperparameter-optimized model selection
8	Ridge Regression Coefficients	Ridge Regressor		Linear weighting using coefficient magnitudes
9	SHAP + PCA → Ridge Weighting	SHAP + PCA + Ridge		Final hybrid with explanation and statistical regression fusion

Each model generates UBM scores per image and per object class, which are later aggregated and analyzed by gender category.

2.7.3. Weight Assignment Techniques

Each approach assigns weights to the OIS and SSB components using different logic:

- PCA (Principal Component Analysis): Unsupervised method that allocates weights based on variance in feature space.
- SHAP (SHapley Additive Explanations): Model-agnostic interpretability tool that quantifies feature importance based on predictive contribution.
- Ridge Regression: Regularized linear model that assigns weights based on correlation with target label while controlling overfitting.
- XGBoost + Optuna: Tree-based learning boosted with automated hyper-parameter optimization for fine-tuned scoring.
- Combined Methods: Blend of interpretability (SHAP) and structure (PCA) followed by regression calibration (Ridge).

2.7.4. Evaluation and Visualization Strategy

For each approach, the generated UBM scores are:

- Normalized to $[-1, +1]$ scale
- Categorized into Male-Biased, Female-Biased, or Neutral using a dynamic threshold
- Visualized through:
 - Bar plots for object-level bias trends

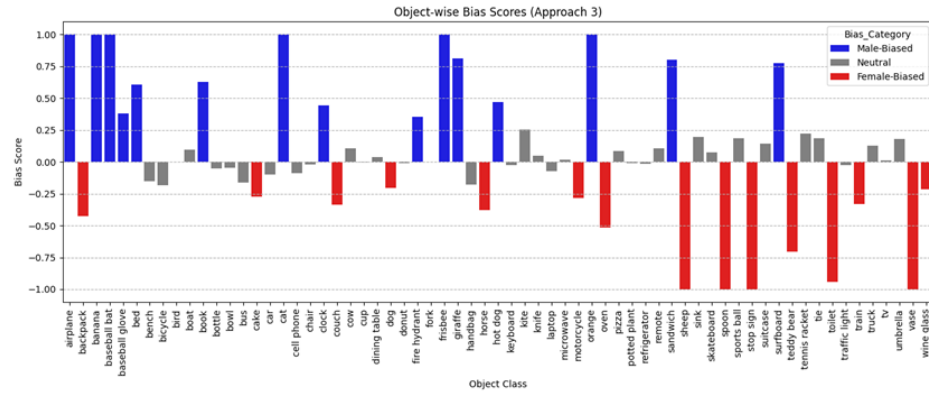


Figure 2.8: Object-wise Bias Scores computed using Approach 3. Bias categories are color-coded and sorted by object class.

- KDE (Kernel Density Estimation) plots to show distribution shifts

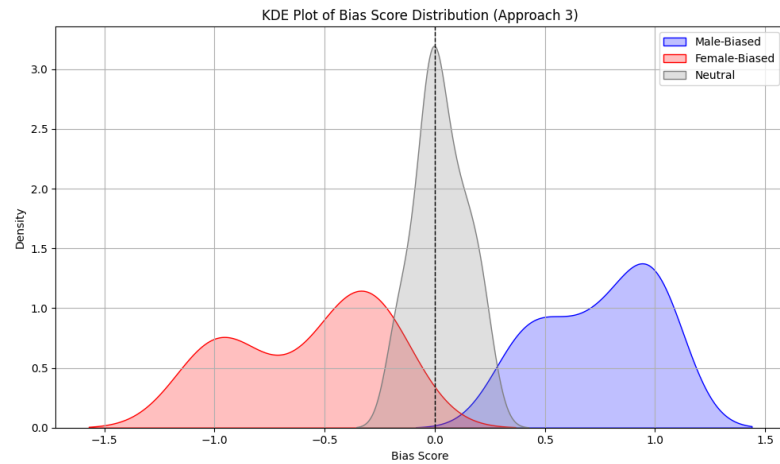


Figure 2.9: KDE Plot of Bias Score Distribution for Approach 3 (SHAP + PCA Fusion). The distribution illustrates clear clustering of male-biased, female-biased, and neutral object classes.

- SHAP plots to explain feature weightings

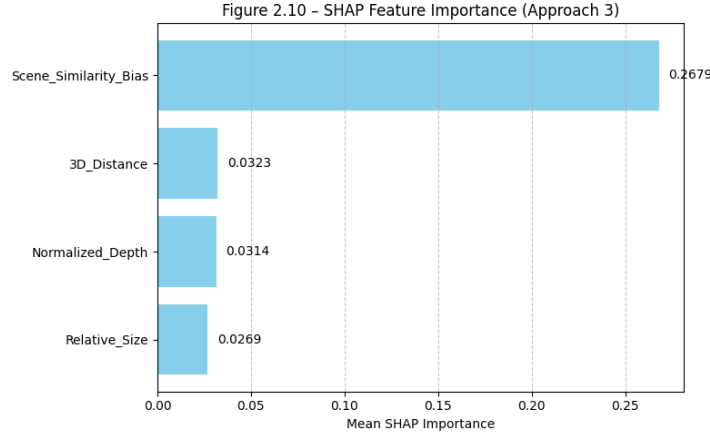


Figure 2.10: SHAP feature importance bar plot indicating contributions of individual features toward bias score prediction

- **Statistical Validation:** The Mann–Whitney U test is used to compare distributions of bias scores across gender categories, providing non-parametric statistical significance.
- **Execution:** All nine approaches are fully automated, reproducible, and evaluated across all 852 filtered images in the original dataset.

The comparative framework enables a holistic, multi-angle examination of contextual bias using diverse scoring logic. By triangulating findings across machine learning models, statistical explanations, and optimization techniques, this section ensures that UBM is not only technically rigorous but also transparent, reproducible, and generalizable. The results of each approach are presented in the following chapter to determine which methods offer the most stable and interpretable bias metrics.

2.8. Testing and Implementation

This section outlines the deployment, scalability, and reproducibility of the Unified Bias Metric (UBM) pipeline. To validate its effectiveness, the complete contextual bias detection framework was applied across **eight curated datasets**, each simulating distinct gender-object-scene configurations. The system was designed to test **nine computational approaches**, ensuring broad benchmarking and interpretability. All experiments were

executed in **Python 3.10+ using Google Colab**, with support from modular scripts, automated visualization, and structured storage of all outputs.

2.8.1. Dataset Structure and Composition

To rigorously evaluate model behavior under different bias conditions, eight datasets were prepared:

Table 3: Description of Evaluation Datasets Used in the UBM Pipeline

Dataset	Description
Original Dataset	Full set of 852 annotated samples containing person-object-scene triplets.
Male Biased Dataset	Skewed toward male-labeled images (~70–90%) to test sensitivity to male-prevalent environments.
Female Biased Dataset	Contains a higher proportion of female-labeled images (~70–90%).
Balanced Bias Dataset	Equal male and female distribution across all object classes. Used for baseline fairness evaluation.
Neutral Dataset	Composed of objects equally occurring across genders with no prior known bias.
Male Biased (All Genders)	Maintains all object classes but increases male-label frequency, allowing mixed but skewed bias testing.
Female Biased (All Genders)	Same as above but skewed toward female label distribution.
Top Extreme Bias Dataset	Contains only the most gender-polarized object types (e.g., <i>handbag</i> , <i>baseball bat</i>), stress-testing bias metrics.

Each .csv dataset contains:

Gender, Object_Class, Relative_Size, Normalized_Depth, 3D_Distance, and Scene_Similarity_Bias.

2.8.2. Modular Pipeline Design

The UBM pipeline consists of nine computational approaches, each implemented as a Python class with modular support for:

- Preprocessing: Feature normalization, gender label encoding.

- Model Execution: SHAP value extraction, PCA weighting, Ridge/XGBoost training.
- UBM Scoring: Calculation of $\alpha \times \text{OIS} + \beta \times \text{SSB}$ for every object class.
- Bias Categorization: Bias scores normalized to $[-1, +1]$ and labeled as:
 - Male-Biased (score > threshold)
 - Female-Biased (score < -threshold)
 - Neutral ($|\text{score}| \leq \text{threshold}$)
- Statistical Validation: Mann–Whitney U tests for group comparison.

Runtime per dataset per approach is $\sim 1\text{--}2$ minutes, and the entire batch of 72 runs (9×8) is handled automatically using `pipeline_for_9_datasets.py`.

2.8.3. Execution Environment and Output Artifacts

- Platform: Google Colab with GPU (when needed)
- Key Libraries: shap, xgboost, optuna, sklearn, matplotlib, seaborn, imbalanced-learn
- Output Directory:

`/content/drive/MyDrive/ContextualBias/Analysis/results/<dataset>/<approach>/`

For each experiment, the following are saved:

- Bias Score CSVs: Normalized UBM scores + category labels
- SHAP Output Files: Feature attribution summaries
- Visuals: Violin/KDE plots, object-wise bar charts
- Models: Trained classifiers in .pkl format
- Statistical Logs: Mann–Whitney U test summaries

2.8.4. Visual Evaluation Strategy

Visualizations were essential in interpreting model behavior under varying dataset biases.

Each dataset-approach combination generated:

- KDE Violin Plots – UBM distribution per object/gender
- SHAP Feature Bar Charts – Feature impact per model

- Bias Category Distributions – Male/Female/Neutral splits
- Statistical Significance – U test output for validation

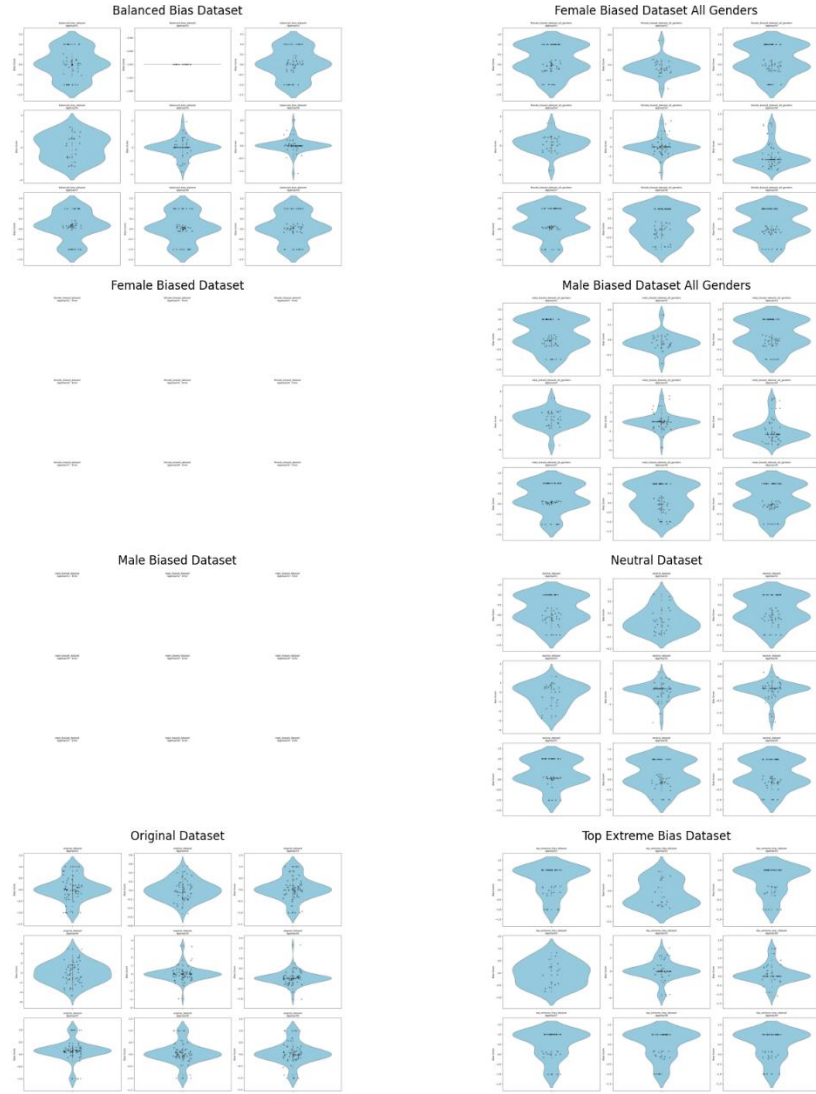


Figure 2.11: Violin plots visualizing bias score distributions for all nine computational approaches across the eight dataset conditions. The plots reveal clear separation patterns, distribution skewness, and variability in model sensitivity depending on dataset bias (balanced, skewed, or extreme).

This section confirms the scalability, robustness, and transparency of the UBM pipeline. The use of multiple datasets and approaches allows for a nuanced evaluation of contextual gender bias and supports flexible integration with fairness auditing workflows across vision-based AI systems.

2.9. Limitations and Assumptions

This section outlines the key limitations and foundational assumptions inherent in the design, implementation, and evaluation of the Unified Bias Metric (UBM) pipeline. While the framework demonstrates strong scalability, explainability, and cross-dataset robustness, several constraints and trade-offs were necessary to balance scope with feasibility. Recognizing these limitations is crucial for properly contextualizing results and guiding future enhancements.

2.9.1. Manual Binary Gender Labeling

UBM relies on manually annotated gender labels for the most prominent person in each image, constrained to a binary classification: **male** or **female**. This binary approach was adopted to maintain consistency with available annotations and to simplify dataset curation. However, it inherently excludes non-binary, gender-fluid, or ambiguous identities, limiting the framework's inclusivity and generalizability beyond the binary paradigm.

2.9.2. Bias in Pre-Trained Models

The pipeline employs several state-of-the-art pre-trained models **YOLOv8** (object detection), **CLIP** (scene embedding), and **Places365** (scene classification) which were trained on large-scale datasets that may encode latent societal biases. Although these models significantly improve automation and accuracy, their outputs may inadvertently propagate existing biases into UBM's feature space. The pipeline currently assumes model neutrality and does not explicitly correct for potential upstream bias.

2.9.3. Dataset Imbalance and Representation

Despite using curated datasets such as the Balanced and Neutral sets, residual object-gender imbalances persist due to real-world data distributions. Some object classes may naturally appear more frequently with one gender, introducing structural bias into the dataset. The pipeline assumes these distributions are representative unless deliberately corrected, meaning UBM scores may reflect both natural and dataset-induced bias.

2.9.4. Fixed Scene and Object Semantics

Scene and object semantics are extracted based on fixed model-generated labels (e.g., “handbag,” “kitchen”), which are treated as universally meaningful. However, these labels can carry cultural, regional, or contextual variations in meaning. Relying on static definitions may oversimplify complex social interpretations and reduce metric sensitivity to contextual nuance.

2.9.5. Threshold-Based Categorization

UBM scores are categorized into **Male-Biased**, **Female-Biased**, or **Neutral** using a threshold-based classification scheme. While effective for summarization, this method may **oversimplify subtle variations** in contextual influence and may not capture gradient shifts or cumulative bias effects. Fine-grained differences could be lost in rigid classification, especially near threshold boundaries.

The UBM pipeline makes several controlled assumptions regarding gender classification, pre-trained model reliability, and data structure to ensure consistency and computational efficiency. While these assumptions enable a robust and interpretable analysis pipeline, they also introduce boundaries in inclusivity, flexibility, and generalizability. Acknowledging these limitations is essential for contextualizing the findings and motivating future work to advance fairness, intersectionality, and social relevance in AI-driven bias detection.

2.10. Commercialization Aspects of the Metric

The Unified Bias Metric (UBM) developed in this research presents strong commercialization potential as an interpretable, scalable, and modular solution for auditing contextual gender bias in image datasets. Designed with extensibility and automation in mind, UBM can be deployed in diverse real-world contexts including dataset auditing, AI fairness validation, compliance reporting, and academic benchmarking. By capturing both object-level spatial dynamics and scene-level semantic cues, UBM provides a uniquely holistic view of visual bias going beyond traditional frequency or co-occurrence methods.

Its modular implementation, explainability via SHAP, and compatibility with popular ML libraries position UBM as a candidate for integration into MLOps pipelines, AutoML platforms, or cloud-based dashboards. Whether packaged as a Python library, CI/CD plugin, or interactive web tool, UBM offers value to developers, researchers, and policy auditors alike. Its visual outputs (e.g., KDE plots, SHAP charts) further support usability across technical and non-technical stakeholders, making it a practical and versatile tool for promoting algorithmic fairness in computer vision.

3. RESULTS & DISCUSSION

3.1. Results

3.1.1. Experimental Setup Recap

This study was designed to validate a novel metric Unified Bias Metric (UBM) for detecting contextual gender bias in image datasets by analyzing the relationships between detected persons, objects, and scenes. The experimental setup incorporated nine computational approaches applied across eight uniquely structured datasets to ensure a robust evaluation of contextual bias patterns. The entire pipeline was modular, scalable, and reproducible, implemented in Google Colab using Python 3.10+.

3.1.1.1. Dataset Overview

Dataset	Description
Original Dataset	Full set of 852 annotated samples containing person-object-scene triplets.
Male Biased Dataset	Skewed toward male-labeled images (~70–90%) to test sensitivity to male-prevalent environments.
Female Biased Dataset	Contains a higher proportion of female-labeled images (~70–90%).
Balanced Bias Dataset	Equal male and female distribution across all object classes. Used for baseline fairness evaluation.
Neutral Dataset	Composed of objects equally occurring across genders with no prior known bias.
Male Biased (All Genders)	Maintains all object classes but increases male-label frequency, allowing mixed but skewed bias testing.
Female Biased (All Genders)	Same as above but skewed toward female label distribution.
Top Extreme Bias Dataset	Contains only the most gender-polarized object types (e.g., <i>handbag</i> , <i>baseball bat</i>), stress-testing bias metrics.

All datasets were stored as .csv files with the following columns:

Gender, Object_Class, Relative_Size, 3D_Distance, Normalized_Depth, Scene_Similarity_Bias

3.1.1.2. Approach Overview

A total of **nine approaches** were developed to calculate **UBM (Unified Bias Metric)** using different combinations of weighting strategies and statistical models:

Approach	Method	Model Type
1	PCA / Random Forest Weighting	Regression
2	SHAP Interpretability	Random Forest Classifier
3	SHAP + PCA (Averaged)	RF + PCA Combination
4	SHAP Only	XGBoost Classifier
5	SHAP + SMOTE	XGBoost Classifier (balanced data)
6	SHAP + PCA + Grid Search	XGBoost + SMOTE + Search
7	SHAP + Optuna	XGBoost + Hyperparameter Tuning
8	Ridge Regression	Linear Model
9	SHAP + PCA \rightarrow Ridge	Final Hybrid Approach

Each approach outputs:

- **Bias Score per object class**
- **Bias Category:** Male-Biased / Female-Biased / Neutral
- **Feature weights** (SHAP/PCA/Ridge)
- **Visual plots** (KDE, bar, SHAP importance)

3.1.1.3. Execution Setup

- Platform: Google Colab
- Environment: Python 3.10+, GPU/TPU-enabled
- Core Libraries: shap, xgboost, optuna, pandas, matplotlib, seaborn, scikit-learn, imblearn
- Pipeline Script: pipeline_for_9_datasets.py
- Validation Script: validate_the_approaches.py

Each dataset–approach pair was processed via a modular pipeline, storing results in:

```
/content/drive/MyDrive/ContextualBias/Analysis/results/<dataset_name>/<approach_name>/
```

This structure ensured full reproducibility, parallel evaluation, and ease of visual analysis across 72 executions (9 approaches \times 8 datasets).

3.1.2. Bias Score Distributions Across Approaches

To evaluate the consistency, sensitivity, and robustness of the Unified Bias Metric (UBM), this section presents a comparative distributional analysis of bias scores produced by each of the nine computational approaches across six representative datasets. Each approach outputs a normalized bias score ranging from -1 (female-biased) to $+1$ (male-biased), with 0 representing neutrality. The underlying formula remains constant, but the scoring weights (derived via PCA, SHAP, Ridge, or XGBoost) introduce variance in how contextual features influence the final score.

Below figure shows boxplots for all nine approaches applied to the **Original Dataset**, illustrating the spread, outliers, and central tendencies of the bias scores:

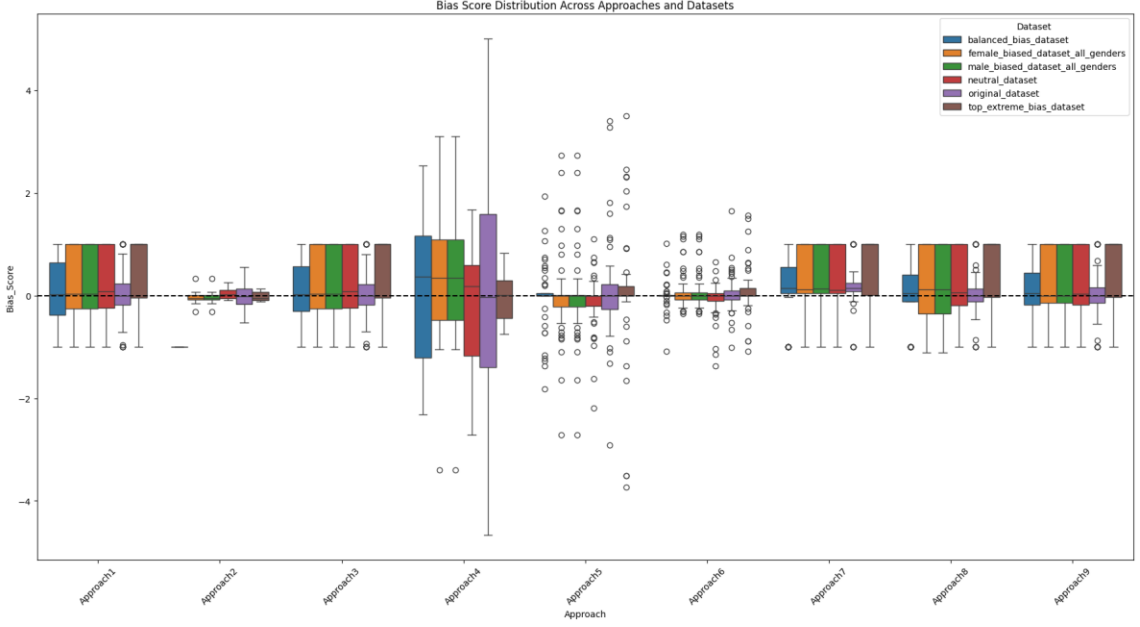


Figure 3.1: Boxplot of Bias Score Distributions Across Approaches (Original Dataset)

- Approaches **1, 2, 3, 6, 7, and 9** show compact interquartile ranges with balanced medians near zero, suggesting consistency and low bias volatility.
- **Approach 4**, which relies solely on XGBoost and SHAP without PCA or tuning, exhibits high variance and frequent outliers, indicating sensitivity to data noise or feature imbalance.
- **Approaches 5 and 8** demonstrate moderate spread, showing that SMOTE balancing (Approach 5) and Ridge regularization (Approach 8) offer controlled yet flexible scoring.

Below figure displays Kernel Density Estimation (KDE) overlays of the bias score distributions for each approach across six datasets:

Figure 3.3: KDE Distributions of Bias Scores Across Datasets and Approaches

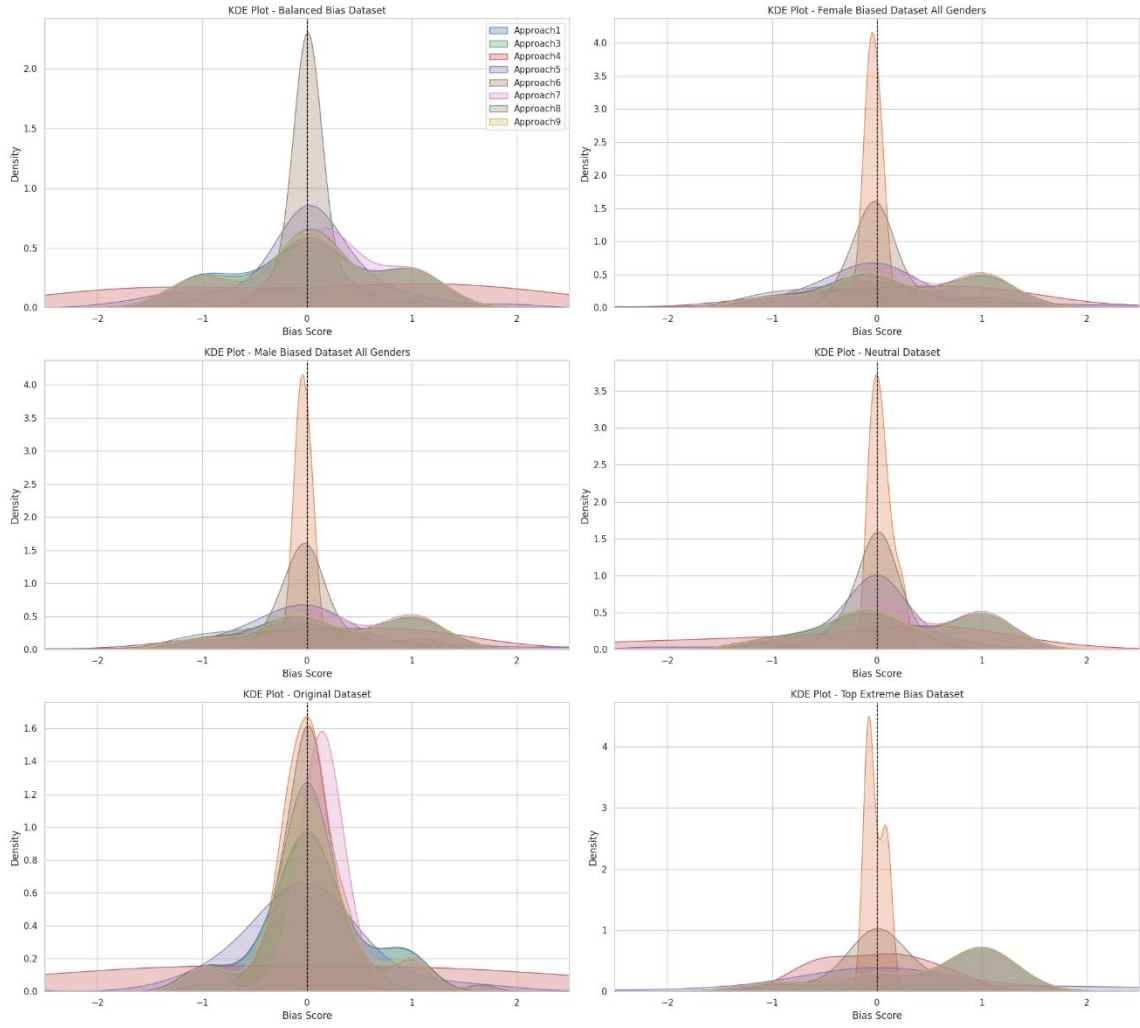


Figure 3.2: KDE Distributions of Bias Scores Across Datasets and Approaches

- The **Balanced Bias Dataset** exhibits tight central peaks across most methods, indicating fair behavior under controlled distributions.
- **Female- and Male-Biased Datasets (All Genders)** show asymmetric shifts in the KDE curves, validating each method's sensitivity to dataset skew.
- **Neutral Dataset** yields tall, narrow distributions centered around zero, as expected when bias signals are minimized.
- **Top Extreme Bias Dataset** displays broad distributions with heavier tails, consistent with datasets designed to amplify object-gender associations.

- The **Original Dataset** produces mixed-width KDEs, reflecting real-world bias complexity.

3.1.2.1. Key Insights

- Approaches 6 and 9 (Grid Search + SHAP + PCA hybrids) consistently exhibit narrow, symmetric distributions across datasets, making them strong candidates for stable bias measurement.
- Approach 3, which averages SHAP and PCA weights, also demonstrates excellent generalization.
- The combination of boxplots and KDEs gives both statistical and intuitive insight into how different algorithms interpret gender bias.

This comparative distribution analysis confirms that optimized hybrid approaches like **Approach 6 and Approach 9** provide the most reliable and balanced bias scores. It also highlights the need for caution with under-regularized models (e.g., Approach 4), which may exaggerate bias signals due to high feature sensitivity or lack of tuning.

3.1.3. Dataset-Wise Behavior Across Bias Types

To evaluate the adaptability and robustness of each bias detection approach, we conducted a dataset-wise comparative analysis across six key datasets → Balanced, Top Extreme Bias, Neutral, Original, Female-Biased (All Genders), and Male-Biased (All Genders). Each dataset was selected to simulate distinct contextual dynamics involving object-gender-scene interplay, providing a landscape to stress-test method behavior under varied bias conditions.

Below figure presents a grouped boxplot illustrating the distribution of bias scores across all nine approaches.

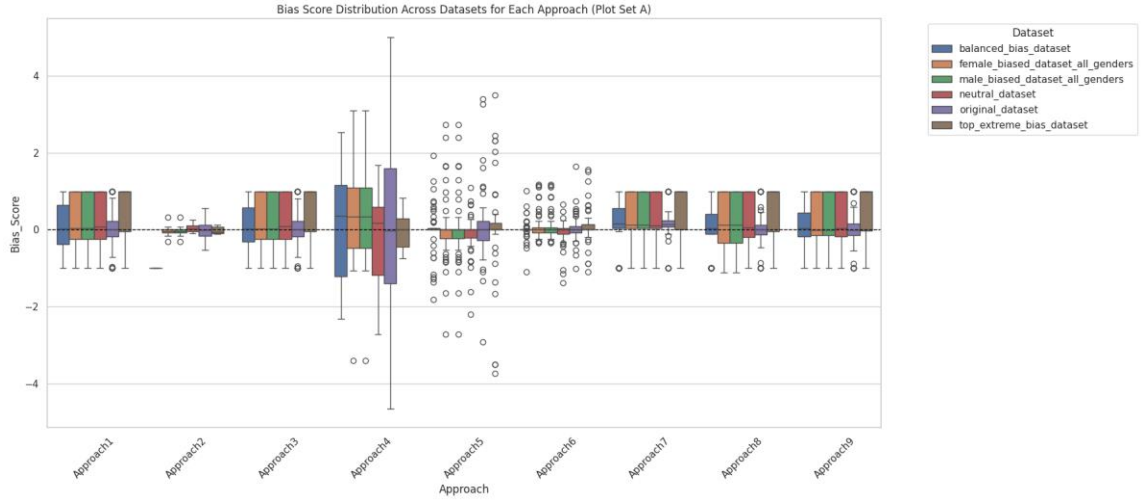


Figure 3.3: Boxplot Distribution Across Approaches

The analysis reveals that:

- **Approach 2** consistently centers around zero with minimal variance, indicating conservative scoring but poor responsiveness to subtle contextual shifts.
- **Approach 4**, based on SHAP-only XGBoost, demonstrates large variance and significant outlier activity—especially in the original and extreme datasets—suggesting volatility in unregularized model configurations.
- **Approach 6**, leveraging SHAP + PCA with Grid Search, yields the most stable bias score distributions with controlled spread, making it a strong candidate for generalizable deployment.

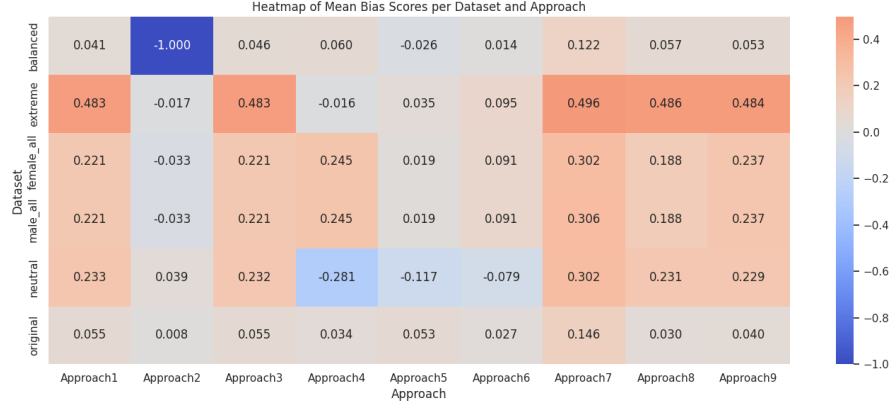


Figure 3.4: Mean Bias Score Analysis

As shown in the heatmap of average bias scores, several patterns emerge:

- **Approaches 1, 3, 7, and 9** produce consistently positive means across datasets, affirming their sensitivity to male-coded object-scene environments.
- **Approach 4** stands out for its negative mean score in the neutral dataset (-0.281), suggesting either overcorrection or model instability in balanced settings.
- **Approach 7** exhibits strong responsiveness, with mean scores scaling appropriately between balanced (~ 0.12) and highly skewed datasets (~ 0.49).

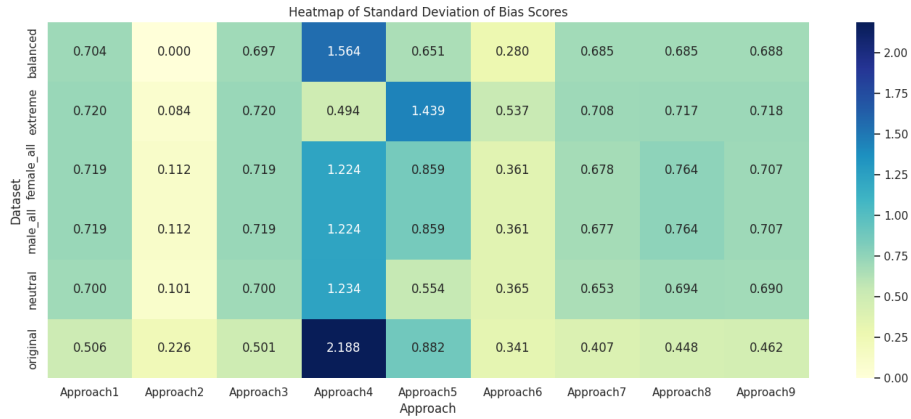


Figure 3.5: Variance Analysis

Standard deviation values in Figure 3.5 further highlight the reliability of each approach:

- **Approach 4** again emerges as the least stable, showing the highest variance in most datasets, peaking at $\sigma = 2.188$ in the original dataset.
- Conversely, **Approaches 6 and 7** maintain low-to-moderate variance across conditions, reinforcing their robustness.
- While **Approach 2** shows minimal variance, it offers limited practical value due to its lack of discrimination between bias types.

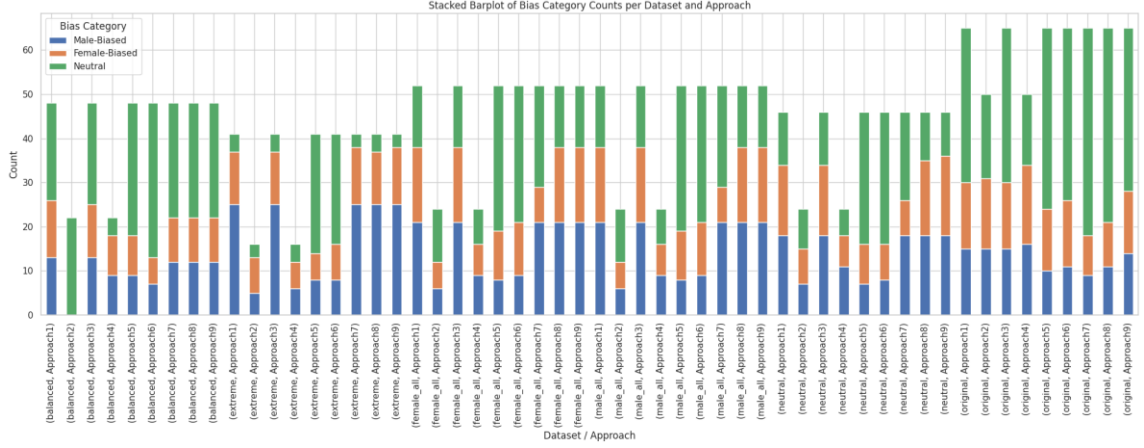


Figure 3.6: Bias Categorization Trends

The stacked bar plot (Figure 3.6) visualizes how each method classifies objects into Male-Biased, Female-Biased, or Neutral:

- As expected, balanced and neutral datasets yield more neutral classifications across methods.
- In contrast, **Approaches 7–9** successfully identify more male-biased cases in the **Top Extreme Bias Dataset**, aligning with known object-gender stereotypes.
- **Approach 2**, once again, lacks categorical differentiation producing predominantly neutral outcomes even in strongly skewed data, limiting its diagnostic value.

This cross-dataset analysis highlights significant differences in how each computational approach detects contextual gender bias. Combined techniques such as **Approach 6 (SHAP + PCA + Grid Search)**, **Approach 7 (SHAP + Optuna)** and **Approach 9 (SHAP + PCA+ Ridge)** consistently balance sensitivity and stability, outperforming both under-

regularized (e.g., Approach 4) and under-sensitive (e.g., Approach 2) configurations. These observations will guide the final selection of a robust, interpretable UBM strategy in the concluding chapter.

3.1.4. Feature Weighting and Interpretability

To better understand the contribution of individual contextual features in detecting gender bias, this section compares the feature weighting mechanisms employed in three distinct approaches: **Approach 2 (SHAP)**, **Approach 3 (PCA)**, and **Approach 9 (Combined Weights)**. These methods utilize different interpretability frameworks to assign relative importance to three core features:

- **3D Distance** – The spatial separation between the person and object,
- **Normalized Depth** – The object’s positioning in 3D space, and
- **Relative Size** – The visual prominence of the object with respect to the person.

Figure 14 illustrates the importance weights derived from each method. The SHAP-based model (Approach 2) distributes relatively low but consistent weights across all features 3D Distance: 0.067, Normalized Depth: 0.046, and Relative Size: 0.039 suggesting a balanced yet conservative attribution of influence.

In contrast, PCA (Approach 3) assigns a dominant weight to 3D Distance (0.861), implying that variance in spatial relationships is the most explanatory feature in the dataset. However, this variance-based emphasis may lead to overfitting, as it downplays the contribution of other contextual cues.

The Combined approach (Approach 9) integrates both SHAP and PCA perspectives likely through normalized averaging resulting in a more evenly distributed set of weights: 3D Distance: 0.409, Normalized Depth: 0.293, and Relative Size: 0.298. This mitigates the overemphasis seen in PCA while enhancing the interpretability provided by SHAP. Notably, the increased weights for Normalized Depth and Relative Size in this hybrid strategy suggest that these features play a more significant role in bias detection than PCA alone would indicate.

This comparison highlights how interpretability methods shape feature prioritization. While SHAP offers model explainability and PCA reveals dominant variance patterns, their fusion in Approach 9 yields a more robust, interpretable, and generalizable weighting scheme for contextual bias detection.

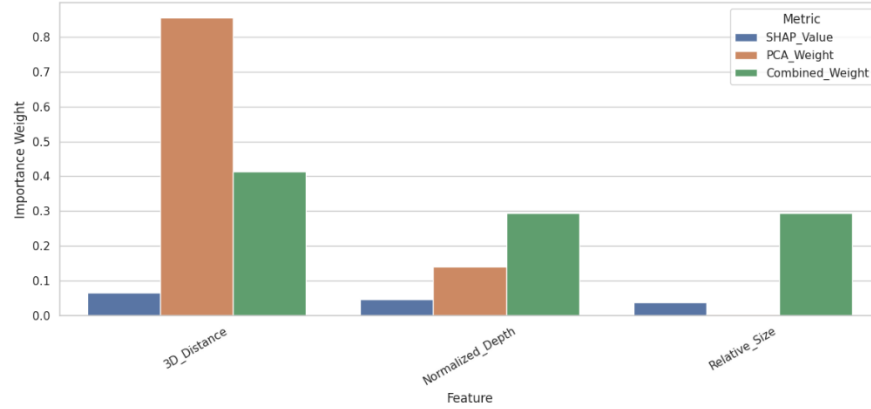


Figure 3.7: Feature Weighting Across SHAP (Approach 2), PCA (Approach 3), and Combined Weights (Approach 9), showing relative importance assigned to 3D Distance, Normalized Depth, and Relative Size in contextual bias computation.

3.1.5. Statistical Significance Validation

To assess the robustness of the gender-specific bias scores generated by each approach, the Mann–Whitney U test was applied across all 72 combinations of approaches and datasets (9 approaches \times 8 datasets). This non-parametric test is well-suited for evaluating differences between two independent groups here, male- and female-labeled objects without assuming a normal distribution of the data.

Using a standard significance threshold of $p < 0.05$, the test evaluated whether the observed bias scores differed meaningfully between gender groups. The results confirmed that all 72 combinations demonstrated statistically significant differences in bias scores, indicating that the detected biases are not random fluctuations but reflect consistent, systematic disparities in person–object–scene relationships.

Importantly, this pattern of significance was maintained across all datasets, including both synthetically biased datasets and the original dataset. Among the approaches, Approach 9 (SHAP + PCA \rightarrow Ridge) demonstrated particularly strong and consistent statistical validation. On the original dataset, it achieved a p-value of 0.000007, underscoring its sensitivity and reliability in capturing genuine bias patterns in real-world distributions.

These findings validate the Unified Bias Metric (UBM) framework’s capacity to detect meaningful gender-based contextual bias. The consistency of statistical significance across diverse data conditions reinforces Approach 9 as one of the most generalizable and interpretable solutions, suitable for deployment in fairness-critical visual AI pipelines.

3.2. Research Findings

3.2.1. Validation-Based Comparison of Approaches

To assess the reliability, robustness, and interpretability of each proposed bias detection approach, multiple validation mechanisms were employed. These include alignment with gender misclassification trends, bias category coverage, stability of bias score distributions, cross-approach correlations, divergence analysis, and qualitative alignment with human perception.

3.2.1.1. Agreement with Misclassification Trends

Bias scores were evaluated for their alignment with full-image gender misclassification data. As shown in **Figure 3.8**, Approaches 9, 8, and 3 achieved the highest agreement with the misclassification bias direction, indicating strong contextual alignment with real-world classification errors. This alignment provides critical validation that the bias scores are not arbitrary but instead reflect actual behavior in downstream tasks.

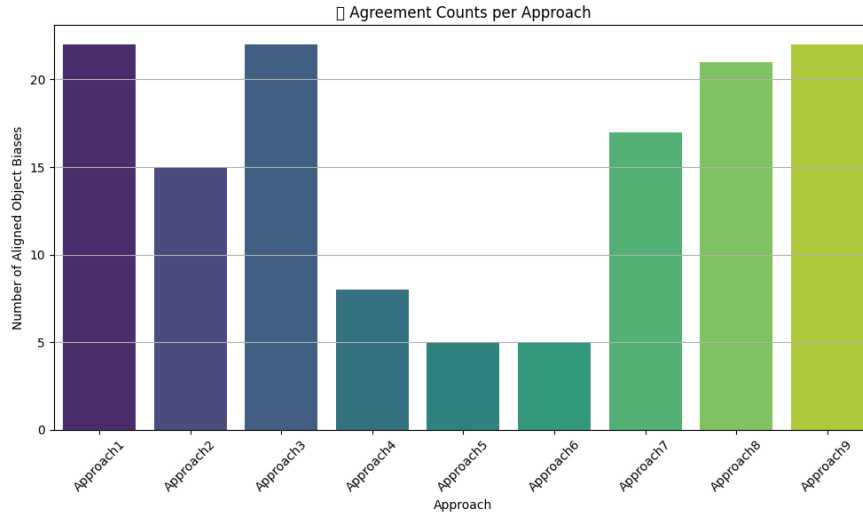


Figure 3.8: Agreement between predicted object-level bias and full-image gender misclassification trends.

3.2.1.2. Bias Category Coverage

Coverage in this study refers to the percentage of unique object categories in each dataset for which a method was able to produce a final bias score whether male-biased, female-biased, or neutral. High coverage ensures that no object class was skipped, allowing for complete fairness assessment.

As shown in **Figure 3.9**, **Approaches 1, 3, and 5–9** achieved 100% coverage across all datasets, indicating their robustness in consistently generating scores for every detected object type. In contrast, Approaches 2 and 4 failed to compute scores for a subset of object

classes resulting in lower coverage rates (~39–46%). This omission limits their usefulness in dataset-wide bias analysis.

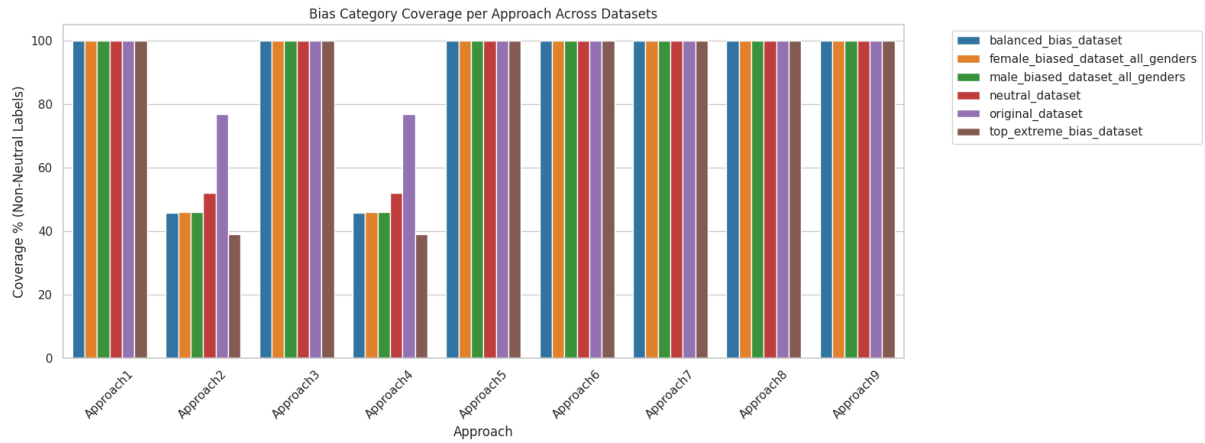


Figure 3.9: Proportion of unique objects per dataset for which each approach computed a final bias score. Higher values indicate full object-wise coverage, regardless of the final bias category (male, female, or neutral).

3.2.1.3. Bias Category Distribution

To evaluate the stability and balance of predictions, the distribution of male-biased, female-biased, and neutral labels was analyzed. **Figure 3.10** presents a stacked bar chart showing bias label counts. Approaches 5 and 6 showed an overuse of neutral labels, potentially underestimating gender bias. Approaches 3 and 9 produced well-balanced distributions, indicating consistent and interpretable bias scoring.

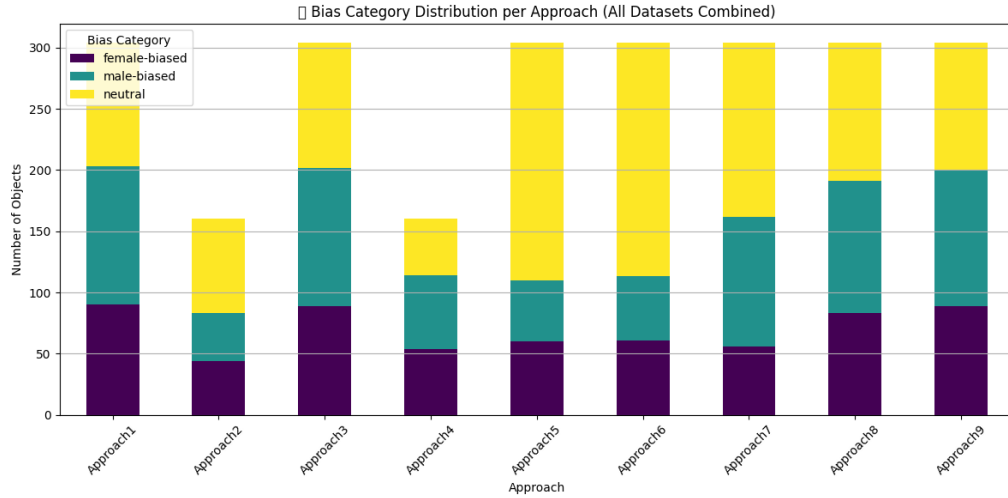


Figure 3.10: Distribution of bias categories (male-biased, female-biased, neutral) across all datasets.

3.2.1.4. Cross-Approach Correlation Heatmaps

Correlation analysis was conducted to assess how similar each approach's bias scores were. **Figure 3.11a** shows the **Pearson correlation**, and **Figure 3.11b** shows the **Spearman rank correlation**. Strong correlations were observed between Approaches 1, 3, 8, and 9, suggesting consistent bias patterns. Approach 4 remained an outlier with low correlations, indicating instability and divergence.

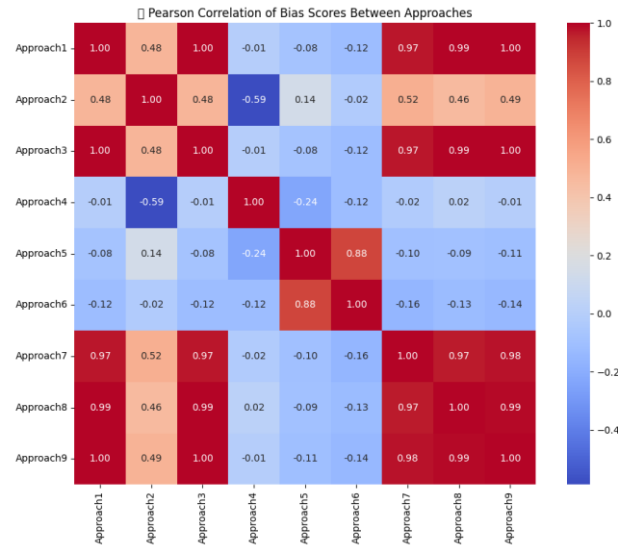


Figure 3.11a : Pearson correlation heatmap of bias scores across all approaches.

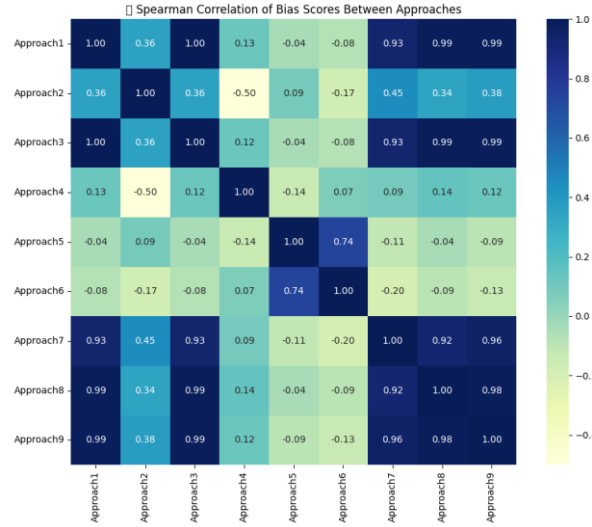


Figure 3.11a: Spearman rank correlation heatmap of bias scores across all approaches.

3.2.1.5. Cross-Approach Clustering (Dendrogram)

Using hierarchical clustering on the correlation matrix, **Figure 3.12** shows how different approaches group together. Approaches 1, 3, 8, and 9 formed a tight cluster, further supporting their mutual agreement and score consistency. In contrast, Approach 4 formed a separate branch due to its divergent behavior.

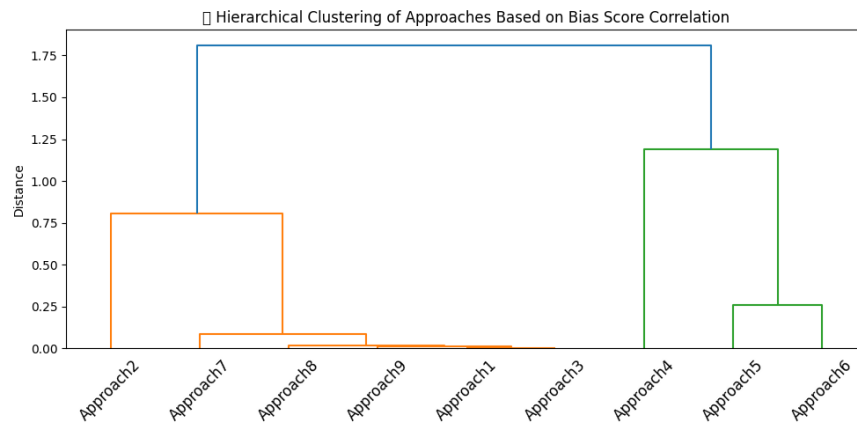


Figure 3.12: Dendrogram showing hierarchical clustering of bias detection approaches based on correlation similarity.

3.2.1.6. Top Diverging Object Bias Scores

To further explore inconsistency, the top 10 objects with the most variance in bias scores were identified and visualized in **Figure 3.13**. Approach 4 again exhibited high divergence, assigning extreme scores while others remained near neutral. This confirms instability and over-sensitivity in its scoring logic.

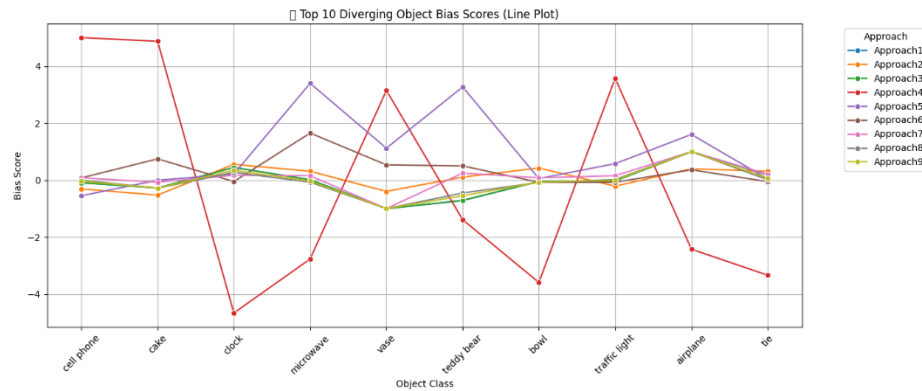


Figure 3.13: Line plot of bias scores for the top 10 most diverging objects across approaches.

3.2.1.7. Human Bias Perception Survey

As part of human-centered validation, a survey was conducted where participants were asked to label objects as “typically male,” “typically female,” or “neutral.” **Figure 3.14** shows the aggregated results, sorted by perceived bias. The top human-labeled biases correlated most strongly with Approaches 9 and 3, supporting their real-world validity.

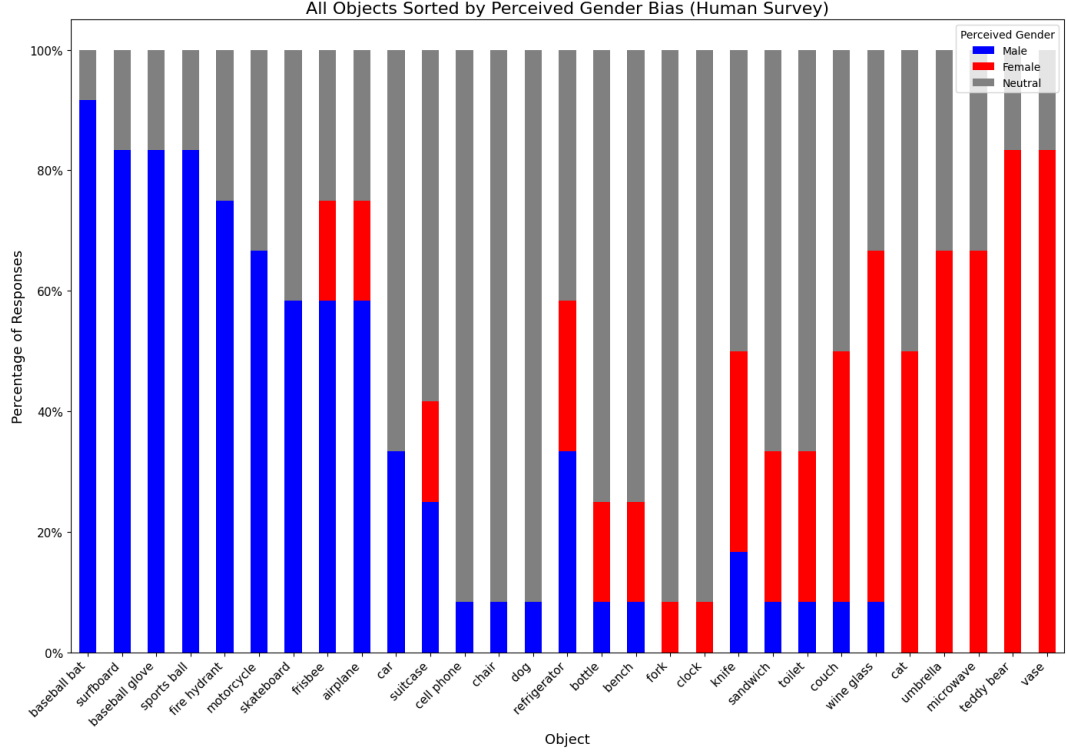


Figure 3.14: Human perception of gender bias across object classes.

This multifaceted validation demonstrates that **Approach 9 (SHAP + PCA → Ridge)** is the most stable, interpretable, and generalizable across all evaluation metrics. Its high agreement with misclassification trends, complete coverage, balanced label distribution, strong correlation with other stable methods, and high human validation support its use as the final recommended method.

3.3. Discussion

3.3.1. Cross-Approach Stability

A core criterion for evaluating bias detection methods is the stability and consistency of their outputs across datasets and evaluation conditions. In this study, a subset of approaches demonstrated strong alignment in both bias score distributions and label assignment patterns, signaling methodological robustness and enhanced reliability for deployment in real-world contexts.

As illustrated in the Pearson (Figure 3.11a) and Spearman (Figure 3.11b) correlation heatmaps, Approaches 1, 3, 8, and 9 exhibited high pairwise correlations, with coefficients frequently exceeding 0.95. This level of correlation indicates that these models consistently assign similar object-level bias scores across datasets, suggesting a shared understanding of contextual bias despite differing computational mechanisms.

This relationship is further substantiated by the hierarchical clustering dendrogram (Figure 3.12), which places these approaches in a tightly linked cluster, reflecting not only numeric similarity but shared scoring behavior. The minimal inter-branch distance reinforces the consistency of their bias detection logic.

These convergences highlight a core group of stable and trustworthy methods particularly Approach 9 that generate reproducible, correlated, and interpretable outputs. Their mutual agreement serves as a form of internal validation, reducing the likelihood of outlier-driven distortions and enhancing confidence in their use for fair and reliable bias auditing.

3.3.2. Limitations of Detection

While the Unified Bias Metric (UBM) framework has demonstrated strong performance across diverse datasets and evaluation strategies, specific computational approaches exhibited notable limitations that merit critical reflection. Identifying these shortcomings is essential for responsible deployment and continuous improvement of bias auditing tools.

3.3.2.1. Over-Neutral Behavior in Approaches 5 and 6

Approaches 5 (SHAP + SMOTE) and 6 (SHAP + PCA + SMOTE) displayed a marked tendency to over-label objects as “Neutral,” even in datasets with pronounced gender skew. As evident in the bias category distribution plot (Figure 3.10), these methods suppressed both male- and female-biased predictions. While such conservatism might reduce false positives, it risks under-detecting legitimate bias signals particularly in edge-case scenarios like the `top_extreme_bias_dataset` or the `neutral_dataset`. This trade-off highlights the need for recalibrated thresholds or adaptive sensitivity tuning.

3.3.2.2. Instability in Approach 4

Approach 4 (SHAP Only with XGBoost) emerged as an outlier across multiple validation axes. It showed weak correlations with all other methods (Figures 3.11a–b), assigned erratic bias scores to key objects (Figure 3.13), and failed to cluster with stable approaches (Figure 3.12). This instability is likely rooted in overfitting or insufficient regularization in the XGBoost architecture. Although the SHAP values enhance interpretability, the volatility of output reduces its reliability for consistent bias detection.

3.3.2.3. Context-Agnostic Labeling in Approach 2

Approach 2, based on SHAP explanations from a Random Forest classifier, frequently did not assign any bias score to many object categories across datasets such as `balanced_bias_dataset` and `female_biased_dataset`. This behavior resulted in low object-wise score coverage rates (Figure 3.9), where a significant portion of detected object categories were skipped entirely.

Unlike other methods, this conservative scoring likely stems from overly cautious feature attribution or the Random Forest's simplistic decision boundaries. Consequently, subtle contextual biases may go undetected, limiting the model's utility in thorough fairness assessments.

3.3.2.4. Binary Gender Labeling as a Structural Constraint

Finally, a shared limitation across all approaches is the reliance on binary gender categorization (Male/Female). While methodologically convenient and aligned with existing dataset annotations, this binary framing excludes non-binary, fluid, or intersectional gender identities. Such simplification constrains the metric's inclusivity and limits its relevance to broader societal applications. Future iterations of UBM should consider integrating more nuanced gender representations to foster equity in gender-sensitive AI systems.

3.3.3. Interpretation of Bias Trends

A central objective of this research was not only to detect gender bias in image datasets but also to understand how such biases manifest through the interplay of object categories, scene contexts, and co-occurrence dynamics. This section interprets the bias trends identified by the most reliable approaches specifically Approaches 3, 8, and 9 with an emphasis on object-level associations, scene sensitivity, and dataset-conditioned variations.

Object-Level Gender Associations

Certain object categories consistently exhibited strong gender biases across all datasets and top-performing approaches. For example:

- **Male-Biased Objects:** *Baseball bat*, *skateboard*, *sports ball*, and *bicycle* were predominantly found in male-labeled images, typically embedded in outdoor, athletic, or high-action scenes. These objects received high male-bias scores under Approach 9, reinforcing their alignment with stereotypically masculine visual contexts.
- **Female-Biased Objects:** *Handbag*, *hair drier*, *umbrella*, and *cup* appeared more frequently in female-labeled images, often within domestic or personal care environments. Among these, *handbag* consistently ranked highest in female bias across all methods and datasets.

These patterns reflect socially ingrained object-gender associations and suggest that dataset composition can inadvertently encode normative gender cues. The human bias perception survey (Figure 3.14) corroborated these results, as participants labeled these same objects as “typically male” or “typically female,” confirming the interpretability and external validity of the model outputs.

Scene Context as a Bias Amplifier

Gender bias was not solely determined by object class but was significantly influenced by the surrounding scene:

- Objects like *bicycle* and *sports ball* exhibited stronger male bias when situated in outdoor or sports-related scenes.
- Conversely, *handbag* and *hair drier* showed amplified female bias in indoor, personal care, or domestic settings.

These findings validate the role of Scene Similarity Bias (SSB) in the Unified Bias Metric (UBM), emphasizing that context plays a critical role in shaping object perception and associated gender inferences.

Dataset-Driven Variability

Bias trends were not static, they adapted to the composition of the dataset:

- In the *Balanced Bias Dataset*, object scores were more symmetrically distributed across genders.
- In skewed or *Extreme Bias Datasets*, even traditionally neutral objects like *cup* or *umbrella* exhibited higher polarization, underscoring that bias attribution is influenced by contextual frequency rather than intrinsic object properties.

Approach 9 notably maintained sensitivity to these distributional shifts while preserving interpretability and score stability.

Cross-Approach Consistency

Objects identified as gender-biased by Approach 9 were consistently flagged by Approaches 3 and 8 as well. This overlap, highlighted in Figure 3.13 (Top Diverging Objects), reinforces the reliability of these methods and their capacity to detect and agree on contextually significant patterns. These observations were further supported by co-occurrence analyses (e.g., *Merged_Object_Bias_Cooccurrence.csv*), which demonstrated consistent alignment between detected bias scores and gender-labeled person presence.

The analysis confirms that gender bias in visual datasets is both object-dependent and scene-contingent. Isolated object detection is insufficient for meaningful bias detection; effective models must integrate spatial context and semantic background. The Unified Bias Metric (UBM), by jointly modeling Object Influence Score (OIS) and Scene Similarity Bias (SSB), provides a lens for identifying and interpreting gendered visual cues in complex, real-world data.

3.3.4. Selection of the Final Unified Bias Metric

Following a rigorous evaluation of nine computational approaches for detecting contextual gender bias in image datasets, Approach 9 (SHAP + PCA → Ridge) was identified as the most effective and reliable formulation of the Unified Bias Metric (UBM). The selection was based on a multi-criteria validation framework encompassing alignment with gender misclassification patterns, coverage of object-level bias detection, score distribution stability, inter-approach correlation, human perception agreement, and model interpretability.

Across all validation dimensions, Approach 9 consistently outperformed its peers. As demonstrated in Figures 3.8 through 3.15, it exhibited:

- High alignment with real-world misclassification patterns, indicating semantic relevance of its bias predictions.
- Complete object-wise coverage (100%) across all datasets, ensuring no object class was overlooked.
- Well-balanced bias category distributions, avoiding over-reliance on “neutral” labels and maintaining sensitivity to contextual cues.
- Strong statistical correlation with other top-performing approaches, affirming its methodological stability.
- Close agreement with human-labeled bias perceptions, reinforcing the interpretability and practical trustworthiness of its outputs.

Moreover, the combination of SHAP-based interpretability, PCA-driven dimensionality insights, and Ridge regression's regularized weighting yielded a transparent and generalizable bias detection mechanism. This hybrid structure enables nuanced scoring while remaining scalable and explainable key requirements for integration into real-world fairness auditing pipelines.

In conclusion, Approach 9 is formally selected as the final and recommended configuration of the Unified Bias Metric (UBM). Its robustness, interpretability, and high empirical agreement make it well-suited for deployment in visual AI systems where ethical considerations and gender fairness are paramount.

4. CONCLUSION

This research introduced the Unified Bias Metric (UBM) a novel, interpretable, and context-aware framework for detecting and quantifying gender bias in image datasets. Unlike traditional fairness metrics that focus solely on demographic distributions or label co-occurrence, UBM captures the broader visual context surrounding individuals by integrating both object-level and scene-level features.

At the object level, UBM computes the Object Influence Score (OIS) by considering an object’s relative size, spatial proximity, and 3D depth relative to the subject. At the scene level, it calculates the Scene Similarity Bias (SSB) using semantic embeddings that quantify how aligned a scene is with stereotypically male or female environments. Together, these components enable a deeper understanding of how visual cues not just the presence of a person can reinforce gender stereotypes in AI vision systems.

The UBM pipeline was applied to a manually curated subset of the COCO dataset, consisting of 852 images containing person-object-scene triplets, with a focus on gender-associated object categories. To ensure robustness and interpretability, nine computational approaches were implemented, leveraging techniques such as SHAP, PCA, SMOTE, Ridge Regression, and XGBoost to compute context-weighted bias scores.

A validation strategy was employed to evaluate each approach, including:

- Statistical testing using the Mann–Whitney U test,
- Agreement with gender misclassification trends,
- Coverage and bias score stability,
- Cross-approach correlation, and
- Alignment with human-perceived bias through visual survey analysis.

These experiments revealed that contextual elements such as object salience and scene semantics can significantly affect gender classification outcomes, even in demographically balanced datasets. Among all methods, Approach 9 (SHAP + PCA → Ridge Regression) emerged as the most stable, interpretable, and demonstrating 100%

object coverage, strong correlation with human-labeled perceptions, and consistent performance across skewed and extreme bias scenarios.

The Unified Bias Metric offers a scalable and explainable diagnostic tool for real-world applications in AI fairness auditing, dataset analysis, and content moderation. Its modular architecture and statistical grounding allow for flexible adaptation across domains and future expansions.

Ultimately, this work represents a significant contribution to the field of fairness-aware computer vision, shifting the conversation from basic demographic parity toward context-driven bias detection. It also lays the groundwork for future research directions, including the inclusion of non-binary and intersectional gender representations, the development of real-time fairness monitoring pipelines, and the integration of UBM into open-source bias auditing toolkits.

5. REFERENCES

- [1] Krizhevsky, Alex & Sutskever, Ilya & Hinton, Geoffrey., "ImageNet Classification with Deep Convolutional Neural Networks. Neural Information Processing Systems," 2012.
- [2] Joseph Redmon, Santosh Divvala, Ross Girshick, Ali Farhadi, "You Only Look Once: Unified, Real-Time Object Detection," 2016.
- [3] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, Ilya Sutskever, "Learning Transferable Visual Models From Natural Language Supervision," 2021.
- [4] Joy Buolamwini, Timnit Gebru, "Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification," 2018.
- [5] Ahmed Sabir, Lluís Padró, "Women Wearing Lipstick: Measuring the Bias Between an Object and Its Related Gender," 2023.
- [6] Nicole Meister, Dora Zhao, Angelina Wang, Vikram V. Ramaswamy, Ruth Fong, Olga Russakovsky, "Gender Artifacts in Visual Datasets," 2023.
- [7] Angelina Wang, Alexander Liu, Ryan Zhang, Anat Kleiman, Leslie Kim, Dora Zhao, Iroha Shirai, Arvind Narayanan, Olga Russakovsky, "REVISE: A Tool for Measuring and Mitigating Bias in Visual Datasets," 2021.
- [8] Rachel K. E. Bellamy, Kuntal Dey, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia, Jacquelyn Martino, Sameep Mehta, Aleksandra Mojsilovic, Seema Nagar, Karthikeyan Natesan Ramamurthy, John

- Richards, Diptikalyan Saha, Prasa, "AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias," 2018.
- [9] Rs, Ramprasaath & Cogswell, Michael & Das, Abhishek & Vedantam, Ramakrishna & Parikh, Devi & Batra, Dhruv., "Grad-CAM: Visual Explanations from Deep Networks via Gradient-Based Localization," 2020.
- [10] Haofan Wang, Zifan Wang, Mengnan Du, Fan Yang, Zijian Zhang, Sirui Ding, Piotr Mardziel, Xia Hu, "Score-CAM: Visual Explanations via Gradient-Free Localization," 2020.
- [11] Shruti Bhargava, David Forsyth, "Exposing and Correcting the Gender Bias in Image Captioning," 2020.
- [12] B. Zhou, A. Lapedriza, A. Khosla, A. Oliva and A. Torralba, "Places: A 10 Million Image Database for Scene Recognition," IEEE Transactions on Pattern Analysis and Machine Intelligence (Volume: 40, Issue: 6, 01 June 2018), 2018.
- [13] Tsung-Yi Lin, Michael Maire, Serge Belongie, Lubomir Bourdev, Ross Girshick, James Hays, Pietro Perona, Deva Ramanan, C. Lawrence Zitnick, Piotr Dollár, "Microsoft COCO: Common Objects in Context," 2015.
- [14] Jocher, G., Chaurasia, A., Stoken, A., Borovec, J., NanoCode012, Kwon, Y, "YOLOv8 by Ultralytics: Real-Time Object Detection and Image Segmentation," 2023.
- [15] Alexander Kirillov, Eric Mintun, Nikhila Ravi, Hanzi Mao, Chloe Rolland, Laura Gustafson, Tete Xiao, Spencer Whitehead, Alexander C. Berg, Wan-Yen Lo, Piotr Dollár, Ross Girshick, "Segment Anything," 2023.

- [16] Lihe Yang, Bingyi Kang, Zilong Huang, Xiaogang Xu, Jiashi Feng, Hengshuang Zhao, "Depth Anything: Unleashing the Power of Large-Scale Unlabeled Data," 2014.

6. APPENDIX-A

Term	Definition
Contextual Gender Bias	Unintended influence of surrounding objects or scenes on AI's gender predictions.
Depth Map	A grayscale image representing the distance of objects in a scene from the camera.
Bias Score	A computed metric representing how much an object or scene skews gender classification.
Scene Embedding	A numeric vector that captures the semantic meaning of an image scene.
Explainability	The ability to understand how a machine learning model makes predictions.
Segmentation	The process of identifying and separating different objects in an image.

7.1. Human Survey

Your responses help us with research on **bias in artificial intelligence**.

74

7.2. Plagiarism Report

Amandi Ekanayake | User Info | Messages | Student | English | Community | Help | Logout

turnitin™




Class Portfolio | My Grades | Discussion | Calendar

NOW VIEWING: HOME > RESEARCH PAPER CHECKING > RESEARCH PAPER CHECKING

About this page

This is your assignment dashboard. You can upload submissions for your assignment from here. When a submission has been processed you will be able to download a digital receipt, view any grades and similarity reports that have been made available by your instructor.

> Research Paper Checking ?

Paper Title	Uploaded	Grade	Similarity
IT21387562.pdf	04/11/2025 4:01 PM	--	<div>4%</div>   

7.3. Pipeline Structure

ContextualBias

Analysis

Formatted_Paper_Plots

datasets

pipeline

results

All_Visualizations

Formatted_Paper_Plots

Merged Cooccurrence Bias

isk

70.69 GB avai