EMERGE - Emergent Multimodal bias Evaluation, Recognition and Guided Elimination

Anonymous ACL submission

Abstract

As AI systems evolve towards increasingly complex multimodal architectures, their performance continues to improve; however, this progression also introduces new challenges. In this paper, we introduce emergent biases as a fundamental ethical challenge—biases that arise not within individual modalities, but through their complex interactions across multiple stages of the pipeline, where fairness assumptions valid in unimodal contexts no longer hold. Motivated by this, we propose EMERGE - Emergent Multimodal bias Evaluation, Recognition, and Guided Elimination, a stage-wise, source-aware procedural fairness framework for detecting emergent biases and enabling targeted mitigation strategies in multimodal AI systems. We also conduct a comprehensive literature review, finding that current fairness interventions largely overlook these emergent phenomena, leaving significant gaps in the state of the art. We further present a systematic taxonomy categorizing various sources of bias across distinct stages of the multimodal pipeline. Finally, we empirically demonstrate our framework using multimodal vision–language datasets and models, highlighting how emergent biases manifest in real-world systems.

1 Introduction

Multimodal artificial intelligence (AI) systems can interpret and integrate heterogeneous data sources by combining modalities such as text, images, audio, and structured data. They offer a promising pathway toward more context-aware, robust, and intelligent systems (Bouchey et al., 2021; Baltrušaitis et al., 2018) in a wide range of applications, from dialogue systems (Awasthi et al., 2024) and hiring processes (Pena et al., 2020) to critical domains such as healthcare (Cai et al., 2019).

A major concern in multimodal models lies in their computational and algorithmic complexity when integrating diverse data types. While technical challenges associated with multimodal stages have been widely studied (Baltrušaitis et al., 2018), the ethical challenges, particularly those related to fairness remain underexplored.

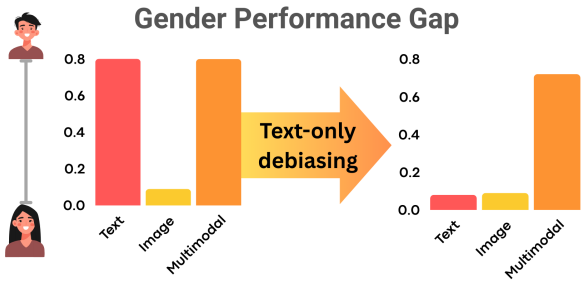

Figure 1: Gender performance gaps associated with text-only, image-only, and multimodal representations before and after applying Counterfactual Data Augmentation to debias text. Although this substantially reduces bias in text-only contexts, it has little impact on gender disparities for multimodal representations.

Fairness failures in text-based NLP systems are extensively studied (Bolukbasi et al., 2016a; Caliskan et al., 2017a). While prior work has shown that biases in text or vision can interact in multimodal architectures (Srinivasan and Bisk), it remains unclear whether these biases are purely compounded from unimodal sources or represent some emergent behaviors, where in the multimodal pipeline they originate, how they evolve across stages, and how to systematically identify the points at which bias dynamics change.

Motivating example In our experiments with the MMBias dataset (Janghorbani and De Melo, 2023), which is distributed under the MIT license, allowing free use, and the CLIP (Radford et al., 2021) model, we first evaluate each modality independently. The text modality exhibits a pronounced true positive rate (TPR) gap, reflecting linguistic biases in the prompts, as shown in Figure 1, whereas the image modality alone remains relatively balanced, since we have nearly the same number of

images per religion and gender. We then use the CLIP model to predict *pleasantness/unpleasantness* using both text and image. The results are unsurprising, as the text modality is severely biased against female categories, showing a TPR difference of 0.8. To address the dominant text modality issue, we apply counterfactual data augmentation (CDA) to the text stream. When we feed only the text modality to predict pleasantness, it achieves a near-ideal fairness score in terms of TPR. As expected, CDA substantially improves fairness in the text modality and narrows its isolated TPR gap. We then combine the de-biased text and image modalities within a shared embedding space using a CLIP-based encoder. However, a substantial parity gap of 0.716 still emerges. The interaction between visual features and linguistic descriptors created a new form of bias that neither the images nor the text independently contains. This is not merely an additive effect: the alignment process transforms and recomposes disparities into a high-magnitude multimodal failure mode, indicating that unimodal interventions alone cannot prevent the reconstruction of inequity in the multimodal pipeline.

To address this, we adopt a procedural fairness lens on the multimodal pipeline, enabling a multifaceted examination of how inequities emerge and evolve. Our key contributions can be summarized as:

(i) We propose EMERGE - a novel stage-wise and source-aware framework based on bias dynamics and procedure fairness for detecting emergent biases;(ii) We present a structured taxonomy categorizing emergent biases and mapping them to distinct stages in the multimodal pipeline, enabling systematic identification and targeted mitigation strategies; (iii) We conduct a systematic literature review focusing on multiple aspects such as bias origin, intervention stage, domain of applications and modalities involved and identified critical gaps; Finally, (iv) we apply our framework to three vision–language datasets (Visogender, MMbias and a curated balanced set) and corresponding models, providing comprehensive validation of our approach and the significance of procedural lens on the pipeline. These datasets have been widely engaged in recent studies (Zhang et al., 2024).

2 Related Works

We conduct a comprehensive review of the current literature concerning fairness in multimodal AI (see appendix A for more detailed and structured domain-wise analysis). To ground the analysis, we begin with a summary of text-debiasing methods in the literature.

Debiasing NLP Early work on bias in NLP revealed that word embeddings encode social stereotypes that propagate into downstream tasks. Bolukbasi et al. (2016b) demonstrated systematic gender associations in popular embeddings and proposed debiasing methods such as hard neutralization and equalization. Subsequent studies showed that contextual models also reproduce and often amplify demographic bias; for example, Zhao et al. (2017) documented gender bias amplification in coreference systems, while Kiritchenko and Mohammad (2018) and De-Arteaga et al. (2019) highlighted racial and gender disparities in sentiment analysis and occupation classification, respectively. Mitigation strategies in NLP have commonly focused on dataset interventions, such as counterfactual data augmentation (Maudslay et al., 2019; Lu et al., 2020) as well as model-level techniques including adversarial training (Zhang et al., 2018) and representation regularization (Liang et al., 2020).

Debiasing in Multimodal Models Multimodal systems (e.g., CLIP, VisualBERT, ViLBERT) introduce additional modes of bias due to the complex interactions between image and text signals. Birhane et al. (2021) and Radford et al. (2021) show that large-scale vision–language models amplify gender, racial, and cultural stereotypes.

Studies have also pointed out that combining multiple modalities can exacerbate risks, rather than eliminate them, especially when the data sources themselves are biased (Agrawal et al., 2018; Alabdulmohsin et al., 2024; Booth et al., 2021).

Efforts to mitigate biases in multimodal learning have been made since 2018, focusing on both the data collection and representation aspects (Pena et al., 2020; Yan et al., 2020; Booth et al., 2021; Alabdulmohsin et al., 2024; Agrawal et al., 2018; Berg et al., 2022). These works primarily involve data balancing, data splitting, and representation learning strategies. Additionally, fusion-based fairness interventions have been explored in (Janghorbani and De Melo, 2023; Kathan et al., 2022; Alasadi et al., 2020). Techniques based on unimodal de-biasing approaches, such as in-processing and post-processing has been studied for mitigating bias in image retrieval (Wang et al.,

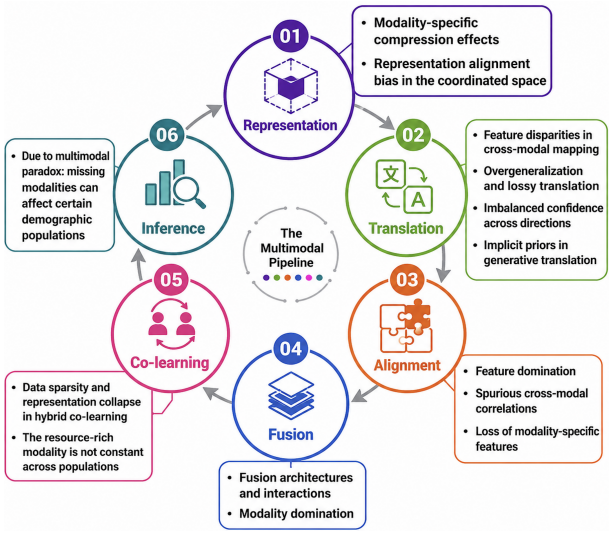

Figure 2: A taxonomy of biases in multimodal AI

2021).

Gap However, current multimodal debiasing work remains fragmented, with each method tied to a specific stage of the pipeline (e.g., dataset curation, representation learning, or caption generation). These approaches use fixed, predefined mitigation strategies that do not adapt to the source or specific stage depends on the bias dynamics. This limitation is critical because biases in multimodal systems are not uniform; they can emerge at stages of the pipeline and interact in complex ways. Therefore, applying blanket mitigation strategies without understanding where and how bias arises may be ineffective or even counterproductive.

3 Bias Sources in Multimodal Pipeline

We first turn our attention to identifying the fairness-related challenges unique to multimodal AI systems. While a taxonomy of core technical challenges is presented in (Baltrušaitis et al., 2018), ethical considerations, particularly fairness, are often overlooked. Building on these foundational issues, we examine each stage of the multimodal pipeline including representation, translation, alignment, fusion, co-learning, and inference, highlighting how cross-modal interactions and specific choices at each stage can introduce new forms of bias, as illustrated in Figure 2.

In the representation stage, multimodal systems learn to encode heterogeneous data into actionable computational formats. Even when individual modalities are balanced and fair, the process of joint or coordinated representation (Zhang et al., 2020) can introduce modality-specific compression or alignment artifacts, causing some modalities (e.g., images) to dominate the representational space. This leads to reduced information retention from other modalities (e.g., text), and such compression can skew predictions across demographic groups.

In the translation stage, converting one modality into another (e.g., images to text) can introduce fairness issues through feature disparities, lossy transformations, or imbalanced reliability across translation directions. These translation models often prioritize prominent features that may not be equally significant across all populations, and can inject implicit generative priors that encode biases absent in the raw data.

At the alignment stage, mapping data between modalities can create spurious associations, such as aligning socioeconomically influenced patterns with specific categories, even when those associations are not justified. Moreover, poor alignment for underrepresented accents or linguistic styles can lead to degraded performance in speech-text models.

Fusion, the process of integrating signals into a unified prediction, introduces further fairness issues as certain modalities may dominate the fusion output based on their availability or predictive strength, reinforcing existing disparities, for example, in healthcare, where imaging or EHR quality may vary by region or population.

In co-learning, where a richer modality supports a weaker one, risks emerge from representation collapse or over-reliance on the dominant modality. Since populations have uneven access to high-quality data sources (e.g., EHRs in urban vs. rural settings), this leads to unfair performance degradation in underrepresented groups.

Finally, at the inference stage, real-world inconsistencies, such as missing or degraded modalities trigger the multimodal paradox: models trained on complete modality sets may behave unfairly when those modalities are absent at test time (Sampath et al., 2025). This issue disproportionately affects groups with lower data quality or access, as the remaining modalities may not adequately compensate.

4 EMERGE: The Framework

Building on our taxonomy of bias sources within the multimodal pipeline, we now introduce a frame-

work to trace such biases. Our framework is grounded in procedural fairness (Wang et al., 2024b), which emphasizes the fairness of the process by which decisions are made, as opposed to focusing solely on fair outcomes (also known as distributive fairness) (Binns, 2018; Oneto and Chiappa, 2020). In multimodal systems, where decision-making involves multiple opaque and interdependent stages, procedural fairness provides a principled lens for tracing how fairness violations arise and persist. Failing to identify new biases that emerge from multimodal interactions can lead researchers to draw fundamentally different or even misleading conclusions.

4.1 Bias Trajectories and Emergent Biases

To measure and track bias dynamics across modalities and stages, we define a contrast function $\Delta(\cdot)$ that quantifies the association gap between a target set (e.g., *pleasantness* vs. *unpleasantness*) and two attribute sets (e.g., *men* vs. *women*). Intuitively, $\Delta(\cdot)$ captures how differently a model associates a target concept with each attribute group, thereby serving as a measure of relative bias.

For each modality, we first encode the target and attribute sets into modality-specific representations. We then compute a per-modality bias contrast using $\Delta(\cdot)$, which reflects the model's behavior within each individual modality. To analyze multimodal interactions, we introduce a fusion function $T(\cdot)$ that combines representations from all modalities into a unified multimodal representation space for both target and attribute sets. We apply $\Delta(\cdot)$ again on this fused representation to obtain the overall bias contrast in the joint space.

Finally, we compare the fused contrast with the unimodal contrasts to characterize the bias trajectory across stages. This comparison reveals whether bias is amplified, attenuated, or preserved after fusion, thereby capturing emergent effects arising from cross-modal interactions. The function $\Delta(\cdot)$ is flexible and can be instantiated using any association metric, such as cosine similarity or caliskan (Caliskan et al., 2017b), making the framework broadly applicable (Appendix B). Here, we do not interpret similarity scores as calibrated probabilities, but as relative metrics to quantify disparities under a common scoring function.

Emergent Bias. We define emergent bias as the setting where unimodal modalities exhibit negligible bias individually, while the fused representation still shows substantial bias. Let δ_{M_i} denote the bias contrast of modality M_i and δ_T denote the bias contrast after fusion via $T(\cdot)$. Emergent bias occurs when unimodal bias remains below a small threshold ϵ, i.e., $|\delta_{M_i}| < \epsilon$, while the fused representation exceeds this threshold and dominates unimodal contrasts, i.e., $|\delta_T| \geq \epsilon$ and $|\delta_T| > |\delta_{M_i}|, \forall i$. Following the 90% rule, we set $\epsilon \leq 0.1$, implying that up to a 10% deviation is considered acceptable.

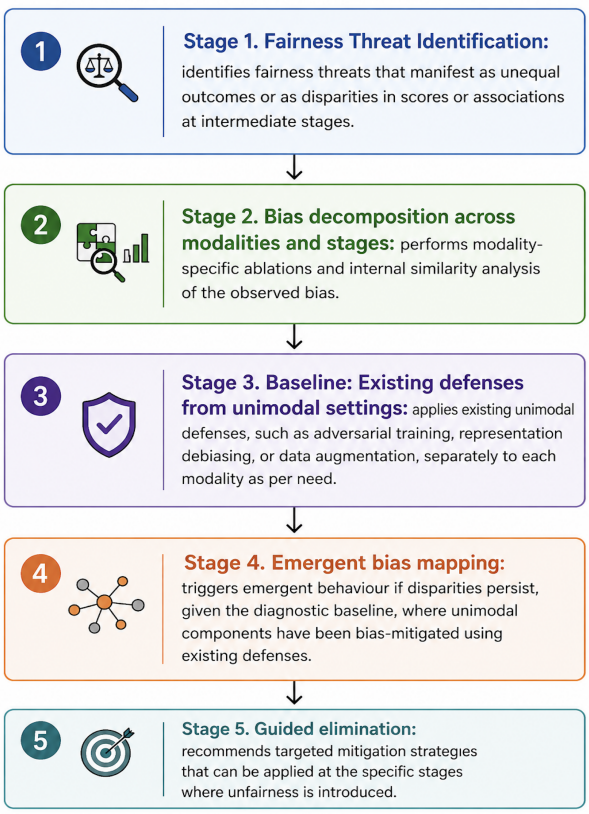

Figure 3: The framework stages

4.2 The Framework Stages

Building on the bias trajectory and emergent bias definition, here, we explain our 5-stage framework in detail as shown in Figure 3.

Stage 1 - Fairness threat identification: The goal of stage 1 is to surface disparities at a specific stage or prediction without prematurely attributing them to specific components within the system.

Stage 2 - Bias decomposition across modalities and stages: Here, we decompose the observed bias by attributing it to individual modalities. To localize its source, we perform modality-specific ablations for text and images, where both the target and attribute sets are confined to the same modality. This allows us to isolate and attribute bias to either the textual or visual component. The result provides insight into each modality's contribution to

disparities. We compute associations directly from the corresponding embeddings, offering a measure of internal representational bias alongside outcome-level disparities.

Stage 3 - Existing defenses from unimodal settings: Once a fairness threat is identified and its modality shares have been quantified, the next stage involves designing or selecting appropriate unimodal mitigation strategies to create a diagnostic baseline. Modalities can be prioritized for intervention based on their isolated bias magnitude. Specifically, we may choose to apply mitigation to (i) all modalities with significant contributions, or (ii) only those exhibiting the lowest representational similarity (e.g., high Caliskan effect size) and greatest outcome imbalance.

Stage 4 - Emergent bias mapping: The persistence of disparities in intermediate (as traced by Algorithm 1) or final outcomes even after mitigation indicates the presence of emergent biases. In such cases, unimodal defenses may prove insufficient, highlighting the need for targeted strategies to address multimodal vulnerabilities.

Stage 5 - Guided elimination: This stage involves creating a checklist of targeted recommendations to address emergent biases. Strategies include selecting fairness-promoting features specifically for biased instances (extending the approach of (Janghorbani and De Melo, 2023)), applying fairness-aware fusion techniques (Alasadi et al., 2020) when multimodal bias surpasses that of individual modalities, and employing causal inference in cases where bias is amplified. By concentrating interventions only where needed, this approach minimizes unnecessary performance trade-offs while effectively enhancing fairness.

5 Empirical Studies

We evaluate EMERGE using CLIP (Radford et al., 2021), and ViLT (Kim et al., 2021) as they expose clearly identifiable stages of multimodal processing. Such architectural transparency is essential for tracing the bias dynamics. CLIP allows us to compute similarity scores between image and text embeddings, which we use to infer sentiment alignment (Esfandiarpoor et al., 2024; Liu et al., 2025). Our goal is therefore not to benchmark state-of-the-art multimodal systems, but rather to study the mechanisms underlying bias dynamics in multimodal architectures.

5.1 MMBias and CLIP

We use the MMBias dataset (Janghorbani and De Melo, 2023) for our preliminary study. Each image is paired with stereotype-relevant text prompts, enabling the creation of multimodal image-text embeddings (See Appendix C for more details). The main prediction task is to classify image-text pairs based on perceived pleasantness or unpleasantness. This classification enables us to detect potential fairness-related threats by observing disparities in predicted sentiments across different identity groups. We use True Positive Rate and Demographic Parity (Pessach and Shmueli, 2022) to assess outcome fairness across groups of interest.

Results and Findings In Table 1, *Fused* represents the multimodal model, specifically, CLIP, evaluated using paired image-text inputs. For **Stage 1**, each pair is classified into pleasant or unpleasant sentiment, and the true positive rate (TPR) is computed to assess disparities across demographic subgroups. Notably, the mental disability and Muslim subcategories exhibit lower TPRs compared to their respective counterparts, suggesting potential biases in the model's sentiment predictions for these groups.

In **Stage 2**, we perform the following: (2a) compute modality-specific TPRs to isolate the contribution of each modality; and (2b) analyse the internal representational similarity gap using Caliskan cosine similarity. In Table 1, *Text* and *Image* represent the corresponding unimodal variants. The text-only modality achieves high TPR for many well-represented groups; however, it underperforms for subgroups such as *mental disability*, *non-disabled*, *heterosexual*, and *American*, indicating that textual signals alone may insufficiently represent these identities. In contrast, the image-only modality performs well for these particular categories, but shows reduced TPRs for groups like *Muslim*, *Jewish*, *Arab*, and *LGBT*, revealing persistent biases in the visual features. Notably, for the subgroups where unimodal performance is lacking in one modality, the complementary information from the other often compensates, resulting in higher TPRs in the *fused* setting. This suggests that integrating multiple modalities can sometimes help mitigate biases inherent in individual ones.

For Stage 2b, we use Caliskan-style cosine similarity and compute the bias contrast directly from the corresponding CLIP embeddings. Figure 4 presents the results of this analysis. Positive val-

Modality	Buddhist	Christian	Hindu	Jewish	Muslim	Mental Dis.	Non-Disabled	Phys. Dis.	Heterosex.	LGBT	American	Arab	Chinese	Mexican
Text	1.00	1.00	1.00	1.00	1.00	0.75	0.85	1.00	0.50	1.00	0.70	1.00	1.00	1.00
Image	0.79	0.76	0.80	0.36	0.25	1.00	0.75	0.78	0.98	0.60	0.92	0.55	0.60	0.55
Fused	0.99	1.00	1.00	0.99	*0.60*	*0.80*	0.95	0.99	0.99	1.00	0.98	0.99	0.99	0.99

Table 1: Stage (2a):TPR comparisons across modalities and subcategories (ideal value = 1) of MMBias

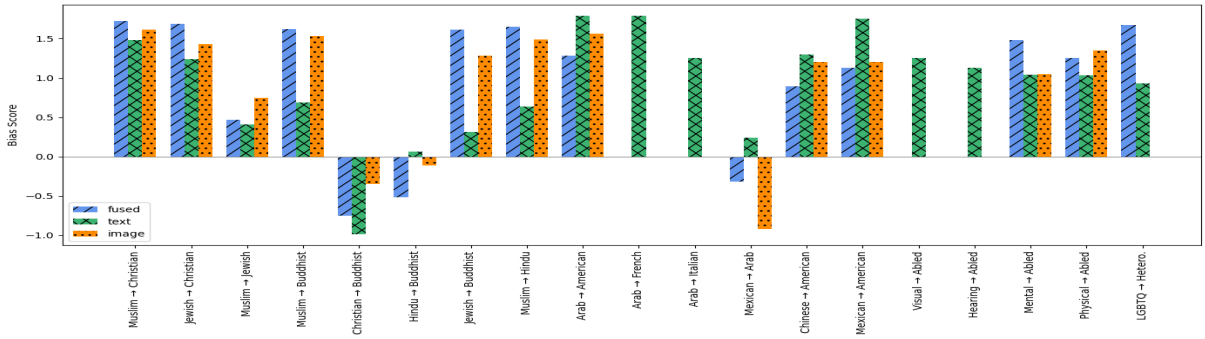

Figure 4: Bias decomposition (2b): Positive values indicate a bias toward the first group listed on the x-axis, while negative values indicate a bias toward the second group.

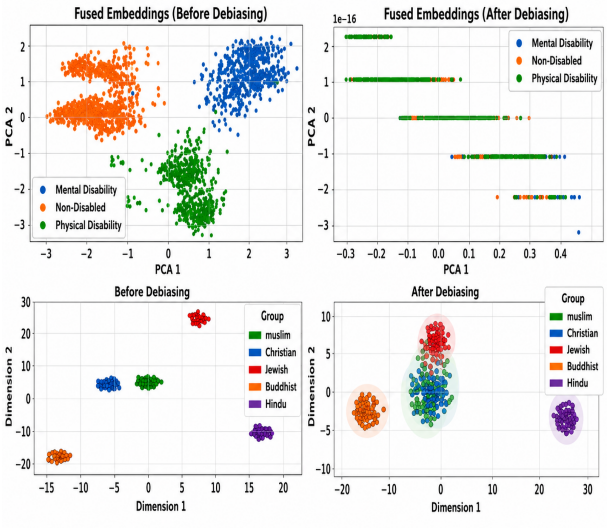

Figure 5: Stage3: Adversarial debiasing for disability (top) and religion (bottom). After debiasing, for disability, the fusion space becomes less separable, while religion remains more clearly distinguishable.

ues indicate a bias toward the first group listed on the x-axis, while negative values indicate a bias toward the second group.We select two categories for subsequent analysis: *disability* and *religion*, where *fused* has higher bias scores than their individual counterparts.

For constructing the baseline (**Stage 3**), we adopt the adversarial debiasing approach from(Zhang et al., 2018), targeting *disability* and *religion* as they exhibit the largest TPR and internal representational disparities in the *fused* model. We ran each experiment for five epochs and computed the average of the results to plot the points shown in Figures 5 and 6.

Figure 5 illustrates the results of adversarial debiasing applied to the *disability* category. This method was independently applied to both image and text embeddings extracted from CLIP and then fused. Prior to debiasing, there is a clear separation in the fused embeddings, based on group membership, indicating the presence of representational bias. Notably, the joint model exhibits signs of bias amplification (Figure 3). After debiasing, these clusters either collapse or significantly overlap, suggesting that group membership is no longer linearly encoded along principal components. This supports the effectiveness of adversarial debiasing in mitigating bias for disability category within the multimodal embedding space. We also analyzed the downstream TPR improvements following debiasing, where the TPR for the mental disability subgroup improved from 0.8 to 0.92 which is in line with prior findings that such procedural transparency as evidenced by the overlapping scenarios enhances outcome fairness (Martin et al., 2022).

For *religion*, the results of debiasing is shown in Figure 5 (bottom). In both before and after debiasing, the fused embedding space remains composed

		Male-Female Similarity	
		Text	Fused
CLIP	OO	0.9304	0.8214
ViLT	OP	0.9790	0.8767
	OO	0.9762	0.7433
	OP	0.9917	0.8068

Table 2: Average male-female similarities for the occupation-object (OO) and occupation-participant (OP) subsets of the Visogender dataset, for both text-only and fused representations. The multimodal representations consistently show lower similarity scores between male and female subjects than text-only representations.

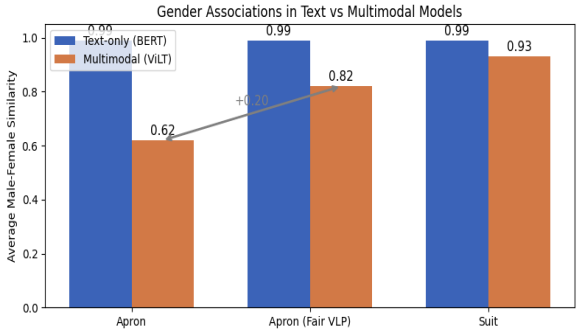

Figure 6: Controlled experiment results: Apron shows strong female association compared to suit in the VILT model; after tailored mitigation using FairVLP debiasing, the gap reduces substantially.

of five separated clusters, one for each religion. This shows that group identity is preserved in the representation, and the debiasing step does not substantially reduce group-wise separability. Also, the TPR for the Muslim subgroup shows a negligible increase from 0.6 to 0.62.

We now proceed to **Stage 4**, where we mark the presence/absence of emergent biases. In the case of the religion, unimodal mitigation proves insufficient, resulting in bias amplification even after debiasing the individual inputs—an indicator of emergent bias. In contrast, for the disability category, unimodal mitigation is effective, with no evidence of bias amplification in either the intermediate representations or the final outcomes (TPR).

Having detected the emergent behaviour, **Stage 5** enables the application of tailored mitigation strategies, rather than indiscriminate interventions. We use the following section (Section 5.2) to show an example of such targeted mitigation, drawn from prior work.

5.2 Controlled Experiment

Here, we first use a curated and balanced dataset in which entities such as *apron* and *suit* are each represented by 50 images of men and 50 images of women, ensuring that any observed bias is not driven by data imbalance. We then used pretrained models to extract representations: BERT (Devlin et al., 2019) encodes text, and ViLT jointly encodes image–caption pairs into multimodal embeddings. For each entity, we compute separate embeddings for male- and female-associated samples and then measures their average cosine similarity, giving an association score that reflects how similarly the model represents the same entity across genders.

Figure 6 shows that apron exhibits a strong gender bias toward females in ViLT (multimodal), while suit maintains similar associations for both males and females in both BERT and ViLT models. Because the images and captions are controlled and balanced, differences in similarity is a fairness threat and captures emergent bias behaviors (as the individual modalities are already balanced).

We conducted additional experiments on the VisoGender dataset (Hall et al., 2023), which is distributed under the Creative Commons license (allowing free use), where we studied the similarities between text-only and multimodal embeddings for male vs. female image-caption pairs. We found that, for both subsets of the VisoGender dataset, Occupation-Object (OO) and Occupation-Participant (OP), male and female text-only embeddings were much more similar than the multimodal embeddings for male vs. female image-caption pairs. The overall results can be found in Table 2. We also isolated objects for which the male-female dissimilarities were particularly striking: *computer*, *stock chart*, *notepad*, and *presentation*. These results can be found in Appendix E.

Targeted mitigation As a targeted remediation, we apply counterfactual augmentation using FairVLP (Zhang et al., 2022) on the joint representation (source of bias) to our first experiment. Our results (*After* in Table 3) show that this approach significantly reduces the observed gender gap, indicating that a substantial portion of the emergent bias arising from cross-modal interactions are mitigated. However, the gap is still not ideal, and the persistence of a gap indicates the presence of residual emergent biases that are embedded within ViLT's internal representations and persist even after applying joint modality debiasing.

			TPR ↑						DP ↑					
			Before			After			Before			After		
			T	I	F	T	I	F	T	I	F	T	I	F
MMBias	CLIP	Muslim	1.00	0.25	0.60	1.00	0.81	0.62*	0.91	0.55	0.75	0.91	0.71	0.76
	CLIP	Christian	1.00	0.76	1.00	1.00	0.91	0.95	1.00	0.81	0.94	1.00	0.86	0.98
	CLIP	Disability	0.75	1.00	0.80	0.86	1.00	0.92	0.71	0.97	0.97	0.80	0.97	0.97
Apron/Suit	ViLT	Gender	0.95	0.61	0.62	0.95	0.76	0.86	0.91	0.54	0.71	0.91	0.78	0.89

Table 3: TPR and DP are reported before and after mitigation across text (T) and image (I) modalities and then fused (F). For MMBias, adversarial debiasing is applied; for Apron/Suit, FairVLP is applied. When both modalities fall below the 90% rule threshold, mitigation is applied to both text and image modalities. The * denotes emergent bias.

Given that both individual data sources and multimodal alignments are bias mitigated at this stage, eliminating bias completely in ViLT may require additional steps such as fairness-aware fusion or explicit regularization. In this way, our proposed framework enables bias identification and mitigation at the point where it emerges, aligning with the principles of a procedural fairness approach that ensures fairness throughout the modeling process rather than applying debiasing indiscriminately at arbitrary stages.

5.3 Summary

In this section, we summarize the empirical analysis in terms of fairness metrics, TPR and DP in Table 3. We examine how fairness differs under unimodal and fused representations, and whether existing defenses at the modality level translate to consistent fairness gains after multimodal stages.

Overall, we observe that unimodal modalities often achieve high or improved scores after mitigation, suggesting that fairness interventions are effective at the single-modality level (for example, *MMBias-Muslim-Image* improved from *0.25* to *0.81* under TPR). However, this improvement does not consistently translate to the fused multimodal representations (negligible improvement from *0.60* to *0.62* for the Muslim category under TPR). This indicates that unimodal improvements do not necessarily guarantee multimodal fairness. We observe that even when individual modalities achieve near-ideal values, the fused outputs can still exhibit noticeable deviation from 1 in both TPR and DP. This shows that there are additional shifts in the intermediate/decision space that are not mitigated by unimodal defenses, reflecting emergent behaviour from various sources as shown in our taxonomy (Section 3; Figure 2).

6 Concluding Remarks

Multimodal AI systems are becoming more prevalent in various critical domains. Therefore, recognizing and addressing emergent biases is essential for ensuring equitable outcomes. Our proposed EMERGE framework systematically maps bias trajectories across multimodal pipelines, revealing how interactions can amplify or create new biases beyond unimodal mitigation efforts.

However, we do not claim that these methods fail universally across all attributes or settings. Instead, our results suggest that the effectiveness of mitigation is highly dependent on the underlying source of bias, and certain attributes require more fine-grained, in-depth analysis guided by our taxonomy to identify the correct stage at which bias originates.

Through empirical analysis on vision-language datasets and models, we demonstrate the practical implications of emergent biases and the limitations of existing fairness techniques. This study underscores the need to move beyond traditional, static fairness definitions towards dynamic, holistic approaches that account for complex cross-modal interactions and targeted source-aware mitigations.

7 Limitations

Our study has certain limitations. Our focus is on detecting and providing targeted mitigation for emergent biases in multimodal systems. Therefore, our evaluation primarily focuses on vision-language datasets and models, which may limit the direct generalizability of findings to other multimodal domains such as audio-text or sensor fusion. Also, while we identify emergent biases and analyze their evolution through the pipeline, we do not propose multimodal-specific de-biasing, instead relying on existing defenses, but targeted mitigation. We leave this as future work.

References

Aishwarya Agrawal, Dhruv Batra, Devi Parikh, and Aniruddha Kembhavi. 2018. Don't just assume; look and answer: Overcoming priors for visual question answering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4971–4980.

Ibrahim Alabdulmohsin, Xiao Wang, Andreas Steiner, Priya Goyal, Alexander D'Amour, and Xiaohua Zhai. 2024. Clip the bias: How useful is balancing data in multimodal learning? *arXiv preprint arXiv:2403.04547*.

Jamal Alasadi, Ramanathan Arunachalam, Pradeep K Atrey, and Vivek K Singh. 2020. A fairness-aware fusion framework for multimodal cyberbullying detection. In *2020 IEEE Sixth International Conference on Multimedia Big Data (BigMM)*, pages 166–173. IEEE.

Varini Awasthi, Rajat Verma, and Namrata Dhanda. 2024. Multimodal dialogue systems in the era of artificial intelligence-assisted industry. In *Machine Vision and Industrial Robotics in Manufacturing*, pages 178–200. CRC Press.

Tadas Baltrušaitis, Chaitanya Ahuja, and Louis-Philippe Morency. 2018. Multimodal machine learning: A survey and taxonomy. *IEEE transactions on pattern analysis and machine intelligence*, 41(2):423–443.

Hugo Berg, Siobhan Mackenzie Hall, Yash Bhalgat, Wonsuk Yang, Hannah Rose Kirk, Aleksandar Shtedritski, and Max Bain. 2022. A prompt array keeps the bias away: Debiasing vision-language models with adversarial learning. *arXiv preprint arXiv:2203.11933*.

Reuben Binns. 2018. Fairness in machine learning: Lessons from political philosophy. In *Conference on fairness, accountability and transparency*, pages 149–159. PMLR.

Abeba Birhane, Dan Card, and David Bear. 2021. Multimodal datasets: Misogyny, pornography, and malicious content. *arXiv preprint arXiv:2106.04312*.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016a. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. In *Advances in Neural Information Processing Systems (NeurIPS)*.

Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016b. Man is to computer programmer as woman is to homemaker? debiasing word embeddings. *Advances in neural information processing systems*, 29.

Brandon M Booth, Louis Hickman, Shree Krishna Subburaj, Louis Tay, Sang Eun Woo, and Sidney K D'Mello. 2021. Bias and fairness in multimodal machine learning: A case study of automated video interviews. In *Proceedings of the 2021 international conference on multimodal interaction*, pages 268–277.

Bettyjo Bouchey, Jill Castek, and John Thygeson. 2021. Multimodal learning. *Innovative Learning Environments in STEM Higher Education: Opportunities, Challenges, and Looking Forward*, pages 35–54.

Qiong Cai, Hao Wang, Zhenmin Li, and Xiao Liu. 2019. A survey on multimodal data-driven smart healthcare systems: approaches and applications. *IEEE Access*, 7:133583–133599.

Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017a. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Aylin Caliskan, Joanna J Bryson, and Arvind Narayanan. 2017b. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

Jiaee Cheong, Sinan Kalkan, and Hatice Gunes. 2024. Fairrefuse: Referee-guided fusion for multi-modal causal fairness in depression detection.

Maria De-Arteaga, Artur Dubrawski, Miroslav Dudik, Sorelle A. Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. 2019. Bias in bios: A case study of semantic representation bias in a high-stakes setting. In *Proceedings of the Conference on Fairness, Accountability, and Transparency (FAT*)*, pages 120–128. ACM.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186.

Reza Esfandiarpoor, Cristina Menghini, and Stephen Bach. 2024. If CLIP could talk: Understanding vision-language model representations through their preferred concept descriptions. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9797–9819, Miami, Florida, USA. Association for Computational Linguistics.

Siobhan Mackenzie Hall, Fernanda Gonçalves Abrantes, Hanwen Zhu, Grace Sodunke, Aleksandar Shtedritski, and Hannah Rose Kirk. 2023. Visogender: A dataset for benchmarking gender bias in image-text pronoun resolution. *Advances in Neural Information Processing Systems*, 36:63687–63723.

Sepehr Janghorbani and Gerard De Melo. 2023. Multimodal bias: Introducing a framework for stereotypical bias assessment beyond gender and race in vision language models. *arXiv preprint arXiv:2303.12734*.

Zifan Jiang, Salman Seyedi, Emily Griner, Ahmed Abbasi, Ali Bahrami Rad, Hyeokhyen Kwon, Robert O Cotes, and Gari D Clifford. 2024. Evaluating and mitigating unfairness in multimodal remote mental health assessments. *PLOS Digital Health*, 3(7):e0000413.

Alexander Kathan, Shahin Amiriparian, Lukas Christ, Andreas Triantafyllopoulos, Niklas Müller, Andreas König, and Björn W Schuller. 2022. A personalised approach to audiovisual humour recognition and its individual-level fairness. In *Proceedings of the 3rd International on Multimodal Sentiment Analysis Workshop and Challenge*, pages 29–36.

Wonjae Kim, Bokyung Son, and Ildoo Kim. 2021. Vilt: Vision-and-language transformer without convolution or region supervision. In *International conference on machine learning*, pages 5583–5594. PMLR.

Svetlana Kiritchenko and Saif M. Mohammad. 2018. Examining gender and race bias in two hundred sentiment analysis systems. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 43–53, New Orleans, Louisiana. Association for Computational Linguistics.

Peizhao Liang, Hao Mu, Xue Wang, and Yang Zhang. 2020. Towards fairness in ai: A survey on bias mitigation in machine learning. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence (AAAI)*, pages 9810–9817. AAAI Press.

Yiming Liu, Yuhui Zhang, Dhruba Ghosh, Ludwig Schmidt, and Serena Yeung-Levy. 2025. Data or language supervision: What makes CLIP better than DINO? In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 1868–1874, Suzhou, China. Association for Computational Linguistics.

Qian Lu, Xiang Li, and Xipeng Qiu. 2020. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 6072–6081. ACL.

Venkata Naresh Mandhala, Debnath Bhattacharyya, and Divya Midhunchakkaravarthy. 2023. A novel study on tools and frameworks for mitigating bias in multimodal datasets. In *Proceedings of the International Conference on Cognitive and Intelligent Computing: ICCIC 2021, Volume 2*, pages 277–283. Springer.

Aaron Martin, Gosia Mikołajczak, and Raymond Orr. 2022. Does process matter? experimental evidence on the effect of procedural fairness on citizens' evaluations of policy outcomes. *International Political Science Review*, 43(1):103–117.

Thomas Maudslay, David Talbot, and A. Black. 2019. Counterfactual data augmentation for mitigating gender bias in text classification. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, pages 4704–4710. ACL.

Luca Oneto and Silvia Chiappa. 2020. Fairness in machine learning. In *Recent trends in learning from data: Tutorials from the inns big data and deep learning conference (innsbddl2019)*, pages 155–196. Springer.

Alejandro Pena, Ignacio Serna, Aythami Morales, and Julian Fierrez. 2020. Bias in multimodal ai: Testbed for fair automatic recruitment. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition Workshops*, pages 28–29.

Dana Pessach and Erez Shmueli. 2022. A review on fairness in machine learning. *ACM Computing Surveys (CSUR)*, 55(3):1–44.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning*, pages 8748–8763. PMLR.

Kishore Sampath, Pratheesh, Ayaazuddin Mohammad, and Resmi Ramachandranpillai. 2025. The multimodal paradox: How added and missing modalities shape bias and performance in multimodal ai. *Preprint*, arXiv:2505.03020.

Matheus Schmitz, Rehan Ahmed, and Jimi Cao. 2022. Bias and fairness on multimodal emotion detection algorithms. *arXiv preprint arXiv:2205.08383*.

Arjun Srinivasan and Yonatan Bisk. Worst of both worlds: Biases compound in pre-trained vision-and-language models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 12345–12356.

Ryan Steed and Aylin Caliskan. 2021. Image representations learned with unsupervised pre-training contain human-like biases. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT)*, pages 701–713. ACM.

Jialu Wang, Yang Liu, and Xin Eric Wang. 2021. Are gender-neutral queries really gender-neutral? mitigating gender bias in image search. *arXiv preprint arXiv:2109.05433*.

Yuqing Wang, Malvika Pillai, Yun Zhao, Catherine Curtin, and Tina Hernandez-Boussard. 2024a. Fairehr-clp: Towards fairness-aware clinical predictions with contrastive learning in multimodal electronic health records. *arXiv preprint arXiv:2402.00955*.

Ziming Wang, Changwu Huang, and Xin Yao. 2024b. Procedural fairness in machine learning. *arXiv preprint arXiv:2404.01877*.

Bo Yan, Wenqi Zeng, Yuqi Sun, Weimin Tan, Xue Zhou, and Chenxi Ma. 2024. The guideline for building fair multimodal medical ai with large vision-language model.

Shen Yan, Di Huang, and Mohammad Soleymani. 2020. Mitigating biases in multimodal personality assessment. In *Proceedings of the 2020 International Conference on Multimodal Interaction*, pages 361–369.

Brian Hu Zhang, Blake Lemoine, and Margaret Mitchell. 2018. Mitigating unwanted biases with adversarial learning. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 335–340.

Chao Zhang, Zichao Yang, Xiaodong He, and Li Deng. 2020. Multimodal intelligence: Representation learning, information fusion, and applications. *IEEE Journal of Selected Topics in Signal Processing*, 14(3):478–493.

Yi Zhang, Junyang Wang, and Jitao Sang. 2022. Counterfactually measuring and eliminating social bias in vision-language pre-training models. In *Proceedings of the 30th ACM International Conference on Multimedia*, pages 4996–5004.

Yunqi Zhang, Songda Li, Chunyuan Deng, Luyi Wang, and Hui Zhao. 2024. Think before you act: A two-stage framework for mitigating gender bias towards vision-language tasks. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 773–791.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. *EMNLP*.

Zhusi Zhong, Jie Li, Shreyas Kulkarni, Helen Zhang, Fayez H Fayad, Yang Li, Scott Collins, Harrison Bai, Sun Ho Ahn, Michael K Atalay, and 1 others. 2024. De-biased disentanglement learning for pulmonary embolism survival prediction on multimodal data. *IEEE Journal of Biomedical and Health Informatics*.

A Literature Review

A.1 Inclusion and Exclusion Criteria

The literature reviewed in this study focuses on peer-reviewed articles, preprints, and conference papers published between 2018 and 2025 that address fairness in multimodal machine learning systems. We specifically target research involving discriminative tasks such as classification, where multiple modalities (e.g., text, image, audio) are used to make predictions. Included studies must explicitly consider issues related to bias, fairness, or equity in these multimodal systems. We exclude studies focused solely on unimodal architectures or that only marginally mention fairness without substantial technical or empirical contributions.

A.2 Search and Selection Strategy

To construct our corpus, we conducted keyword-based searches using combinations such as "multimodal fairness," "bias in multimodal learning," "fair representation multimodal," and "multimodal classification bias" across major digital libraries including Google Scholar, ACL Anthology, arXiv, IEEE Xplore, and ACM Digital Library. We manually screened titles and abstracts for relevance and reviewed full texts where necessary to determine fit with our selection criteria. Forward and backward citation tracking was employed to identify additional relevant studies. Based on this process, we selected a total of 17 publications that met our inclusion criteria and offered substantive contributions to the understanding or mitigation of bias in multimodal machine learning systems.

A.3 Categorization and Annotation Process

Each selected paper was annotated for the domain of application (e.g., general-purpose or healthcare), the type of modalities used, the identified sources of bias (e.g., dataset imbalance, representational skew), and the intervention stages addressed (e.g., pre-processing, representation learning, fusion, post-processing). Mitigation strategies were categorized as either unimodal transfer methods, multimodal-specific interventions, or fused representation debiasing. The annotated information was synthesized to identify prevailing trends and gaps in the literature.

A.4 Review Scope: Exclusion of Generative Models

While recent advancements in generative models (e.g., diffusion models, large multimodal foundation models) have opened new frontiers in AI research, we chose to exclude them from our review for two main reasons. First, fairness concerns in generative settings—such as representation harms, stereotypical outputs, or content filtering biases—are fundamentally different from those in discriminative tasks like classification, which are our primary focus. Second, the evaluation of fairness in generative multimodal systems remains underdeveloped, with limited reproducible metrics or established mitigation pipelines. To ensure conceptual coherence and maintain rigorous comparison across studies, we therefore restrict our scope to discriminative models where multimodal inputs are used for supervised prediction and fairness can be

more concretely assessed.

That said, as generative multimodal systems continue to proliferate and their societal impact deepens, we recognize the importance of extending fairness analyses to these models. In future work, we plan to expand our framework to incorporate generative settings, particularly focusing on evaluation criteria, representational harms, and emergent biases arising from cross-modal generation and alignment.

Fairness in Multimodal AI-General Table 1 provides a summary of the literature on fairness and multimodal AI, focusing on problems identified, solutions proposed, modalities considered, and the corresponding domain of applications.

Fairness in Multimodal AI- Healthcare Table 2 summarizes the literature on fairness in multimodal healthcare applications.

Studies focus on the issue of demographic representation within multimodal datasets (Wang et al., 2024a; Yan et al., 2024). The issue is further exacerbated by the integration of diverse data sources, which, may unintentionally amplify biases against certain groups due to indiscriminate inclusion of patient metadata (Yan et al., 2024). This unbalanced representation of gender, race, and other demographic factors in the data can result in skewed diagnostics and healthcare recommendations (Yan et al., 2024). Evaluating fairness across different modalities, such as image, text, and audio, remains a significant challenge. Jiang et al. (Jiang et al., 2024) compare the fairness of various modalities and in combination. Zhong et al. (Zhong et al., 2024) emphasize the lack of a holistic framework to de-bias multimodal data and propose a solution involving de-biased disentanglement in the pre-trained space. Similarly, Wang et al. (Wang et al., 2024a) argue that existing unimodal fairness methods fail to address the multi-faceted social biases that are often intertwined with demographic factors in multimodal EHRs. All of the above strategies focused at mitigation strategies centered on data. A fusion based strategy is proposed by (Cheong et al., 2024) for depression detection while a post-hoc classifier optimization is proposed for mental health (Jiang et al., 2024).

B Algorithm

We begin by modelling bias trajectory that systematically traces how bias evolves across stages of a multimodal pipeline. Our method estimates bias at each step and categorizes its progression as amplification, attenuation, or preservation, as formalized in Algorithm 1.

The algorithm takes as input multiple modalities along with corresponding target (such as Muslim and Christian) and attribute sets (sentiments), a similarity function, and a transformation function. Using these inputs, it computes association scores for each individual modality as well as for their

Algorithm 1 Bias Trajectory Analysis

Require: Modalities $\mathcal{M} = \{M_1, M_2, \ldots, M_N\}$, target sets X, Y, attribute sets A, B, similarity function $\mathcal{S}(\cdot,\cdot)$, transformation $T(\cdot)$
Ensure: Bias contrasts $\{\delta_{M_i}\}$, fused contrast δ_T, $\Delta\delta_{\max}, \Delta\delta_{\min}$, bias trajectory label

function $\Delta(Z_X, Z_A, Z_B)$
  return $\text{mean}\big(\mathcal{S}(Z_X, Z_A)\big) - \text{mean}\big(\mathcal{S}(Z_X, Z_B)\big)$
end function
for $i = 1$ to N **do**
  Encode X, Y, A, B into modality-specific representations: $Z_{X_i}, Z_{Y_i}, Z_{A_i}, Z_{B_i}$
  $\delta_{X_i} \leftarrow \Delta(Z_{X_i}, Z_{A_i}, Z_{B_i})$
  $\delta_{Y_i} \leftarrow \Delta(Z_{Y_i}, Z_{A_i}, Z_{B_i})$
  $\delta_{M_i} \leftarrow \delta_{X_i} - \delta_{Y_i}$ ▷ Bias contrast per modality
end for
$Z_X^T \leftarrow T(Z_{X_1}, \ldots, Z_{X_N})$ ▷ Fuse target set X
$Z_Y^T \leftarrow T(Z_{Y_1}, \ldots, Z_{Y_N})$ ▷ Fuse target set Y
$Z_A^T \leftarrow T(Z_{A_1}, \ldots, Z_{A_N})$ ▷ Fuse attribute set A
$Z_B^T \leftarrow T(Z_{B_1}, \ldots, Z_{B_N})$ ▷ Fuse attribute set B
$\delta_X^T \leftarrow \Delta(Z_X^T, Z_A^T, Z_B^T)$
$\delta_Y^T \leftarrow \Delta(Z_Y^T, Z_A^T, Z_B^T)$
$\delta_T \leftarrow \delta_X^T - \delta_Y^T$ ▷ Fused bias contrast
$\delta_{\max} \leftarrow \max_i \delta_{M_i}$
$\delta_{\min} \leftarrow \min_i \delta_{M_i}$
$\Delta\delta_{\max} \leftarrow \delta_T - \delta_{\max}$
$\Delta\delta_{\min} \leftarrow \delta_T - \delta_{\min}$
if $\Delta\delta_{\max} > 0$ **then**
  Trajectory ← Bias Amplification
else if $\Delta\delta_{\min} < 0$ **then**
  Trajectory ← Bias Attenuation
else
  Trajectory ← Bias Preservation
end if
return $\{\delta_{M_i}\}, \delta_T, \Delta\delta_{\max}, \Delta\delta_{\min}$, Trajectory

Algorithm 2 Similarity / Association Function

Require: Two sets of embeddings Z_X, Z_Y, similarity method, method $\in$ {cosine, effect_size}
Ensure: Similarity / association score between Z_X and Z_Y

if method = cosine **then**
  Compute pairwise cosine similarities: $c_{ij} \leftarrow \frac{Z_X[i] \cdot Z_Y[j]}{\|Z_X[i]\| \|Z_Y[j]\|}$ for all i, j
  score $\leftarrow \text{mean}(c_{ij})$ ▷ Mean cosine similarity
else if method = effect_size **then**
  Compute mean embeddings: $\mu_X \leftarrow \text{mean}(Z_X)$, $\mu_Y \leftarrow \text{mean}(Z_Y)$
  Compute pooled standard deviation: $\sigma \leftarrow \sqrt{\frac{\text{var}(Z_X)+\text{var}(Z_Y)}{2}}$
  score $\leftarrow \frac{\mu_X - \mu_Y}{\sigma}$ ▷ Cohen's d style effect-size
end if
return score

Table 4: Summary of literature on fairness and multimodal AI – General.

Ref.	Problem	Solution/Findings	Modalities	Domain
(Pena et al., 2020)	Generate feature embeddings from training dataset that are agnostic to gender and ethnicity	Feature embeddings generation	Image, text	Recruitment
(Yan et al., 2020)	Different modalities show various patterns of unfairness, and data fusion introduces emergent risks (gender, race)	Data balancing, adversarial learning	Audio, visual	Recruitment
(Booth et al., 2021)	Combining multiple modalities only marginally improves accuracy but increases unfairness	Gender-norming and reduced feature selection from training data	Text, video	Recruitment
(Janghorbani and De Melo, 2023)	Lack of benchmarks to evaluate unfairness in vision-language models for minorities:religion, nationality, sexual orientation, and disabilities	MM-bias dataset, post-processing	Text, image	Vision-Language
(Alabdulmohsin et al., 2024)	CLIP models inadvertently learn and perpetuate societal unfairness such as representation and association biases from multimodal data	M4 algorithm for data balancing; mixed impacts on performance and effectiveness	Image, text	Vision-Language
(Berg et al., 2022)	Lack of measurement robustness and feature degradation in multimodal fairness	Adversarial Debiasing and contrastive loss for unbiased representation learning	Text, Image	Vision-Language
(Wang et al., 2021)	Study a unique gender bias in image search – the search images are often gender-imbalanced for gender-neutral natural language queries	In-process fair sampling and post-process adjustment	Text, Image	Vision-Language
(Schmitz et al., 2022)	How multimodal data affect fairness and performance	Text alone contributes to majority of performance with least bias	Audio, Text, Video	Emotion Detection
(Mandhala et al., 2023)	Survey on open source frameworks for fairness in data	No solution proposed; not focused on multimodal datasets	N/A	Review
(Kathan et al., 2022)	Lack of personalized models	Multimodal humour recognition personalized per individual using late fusion; no mitigation	Audio, Visual	Emotion (Humour Recognition)
(Alasadi et al., 2020)	Ensure performance equity across demographics in cyber-bullying detection	Fairness-aware weighted fusion to balance accuracy and fairness	Visual, Text	Cyber-bullying Detection
(Agrawal et al., 2018)	Spurious correlations between image and text	Grounded VQA (GVQA) and fair dataset splitting	Text, Visual	VQA

fused combination. It then measures the change in bias across stages and assigns a bias trajectory label that describes whether the bias is amplified, attenuated, or preserved throughout the multimodal pipeline. The similarity function $\mathcal{S}$ can be derived from Caliskan cosine similarity (Caliskan et al., 2017b) or cosine or any association-based scores to measure how strongly input representations (e.g., text, image, or fused representations) are associated with specific attribute concepts (e.g., *pleasant* vs. *unpleasant*) (Janghorbani and De Melo, 2023).

This formulation highlights where bias emerges in the pipeline, allowing targeted interventions rather than relying solely on post hoc corrections. This approach is crucial because bias does not always evolve monotonically throughout the pipeline; some biases may be amplified at certain stages, while others naturally attenuate or remain preserved without requiring intervention. By performing stage-wise analysis, we can accurately identify when and where bias amplification occurs, allowing targeted and effective mitigation only when necessary. At the same time, recognizing natural bias attenuation helps avoid unnecessary corrective measures that could harm model performance or distort important modality-specific information.

C Empirical Study:More Details

Dataset Most of the studies based on multimodal fairness is with gender and racial disparities. The lack of standardised benchmark dataset hinder the development and analysis of fairness issues concerning other groups. To resolve this gap, (Janghorbani and De Melo, 2023) proposed an MMBias dataset which has a broad general inclusion of other categories containing image and text. The details are as follows:

The dataset consists of 3,500 target images and 350 English phrases, each associated with distinct target concepts. Every concept category includes 250 images, sourced from the widely used image-sharing platform *Flickr*. Furthermore, for each target concept, the dataset provides 20 related textual phrases, which are utilized in experiments investigating bias in the textual domain.

The MMBias dataset also includes a valence dataset, composed of two distinct image sets representing visually pleasant and unpleasant concepts. These sets were constructed using a methodology similar to that of Steed and Caliskan (Steed and Caliskan, 2021), by retrieving images associated with positively valenced themes such as *wealth*, *peace*, *babies*, *love*, and *butterflies*, as well as

Table 5: Summary of literature on fairness and multimodal AI in healthcare.

Ref.	Problem	Solution/Findings	Modalities	Demographics
(Wang et al., 2024a)	Unimodal fairness fails to address the multifaceted social biases intertwined with demographic factors in multimodal EHRs	Synthetic data generation with contrastive learning	Longitudinal data, clinical notes	Gender, race, age SES
(Jiang et al., 2024)	Compare fairness of modalities in combination; no statistically significant unfairness was found in the composition of the dataset	Post-training classifier threshold optimization	Image, audio, cardiovascular patterns	Gender, race, educa tion, age
(Cheong et al., 2024)	Lack of work in fairness and multimodal Mental Health	A causal multimodal framework with two modules: one for achieving group fairness through causal debiasing, and another for ensuring individual fairness by adaptively fusing modalities using a fairness-guided mechanism	Audio, text, visual	Gender
(Zhong et al., 2024)	Lack of holistic framework to de-bias multimodal data	De-biased disentanglement in the pre-trained space	Image, text	Race, ethnicity, gen der
(Yan et al., 2024)	Integrating diverse data sources for better diagnostics unintentionally amplifies biases against certain groups due to indiscriminate inclusion of patient metadata	Fairness-oriented strategies for selecting patient metadata, guided by large-scale medical vision-language models (VLMs)	Image, text	Gender, age, skin tone

Target Groups	Subgroups
Religion	Islam, Christianity, Judaism, Buddhism, Hinduism
Nationality	American, Arab, Chinese, Mexican
Disability	Physical disability, Mental disability, No disability
Sexual Orientation	Homosexual, Heterosexual

Table 6: Groups and subgroups in MMbias dataset

negatively valenced themes such as *death*, *injury*, *prison*, and *fear*.

Main task The main task in the MMBias framework is to perform binary classification of inputs:text, images, or their fused representations, based on valence, distinguishing whether a given input reflects a pleasant or unpleasant concept. This task leverages the valence subset of the MMBias dataset, which includes images and phrases associated with emotionally positive concepts (e.g., peace, love, wealth, babies) and negative ones (e.g., death, injury, prison, fear). The classification task serves as the foundation for studying representational biases across modalities and is central to evaluating how well debiasing techniques preserve task-relevant information while mitigating sensitive attribute associations.

CLIP Model We employ CLIP (Contrastive Language–Image Pretraining), a multimodal vision-language model developed by OpenAI (Radford et al., 2021), as the backbone for extracting image and text embeddings in our experiments. CLIP is trained using a self-supervised contrastive learning objective that aligns image and text representations in a shared embedding space. This training leverages a large-scale dataset of 400 million image–text pairs sourced from the web. While originally designed for zero-shot image classification, CLIP has demonstrated strong generalization across a range of downstream tasks, including semantic similarity and retrieval. The architecture comprises separate visual and textual encoders, enabling independent access to each modality's embedding. This modularity is particularly valuable in our setup, allowing us to isolate and analyze bias within and across modalities. For all experiments, we use the "ViT-B/32" variant and rely on the official CLIP implementation released by OpenAI [1]. All experiments are conducted in inference mode only. Running inference over the MM-Bias dataset with standard batching required approximately 4 GPU-hours for CLIP on a single NVIDIA A100-equivalent GPU.

Adversarial Model We implement an adversarial debiasing framework using CLIP ViT-B/32 (OpenAI) as the base model for generating both text and image embeddings, each of dimension 512. A binary logistic regression classifier, opti-

[1] https://github.com/openai/CLIP

mized with the `lbfgs` solver and a maximum of 1000 iterations, is trained to predict target labels (e.g., sentiment: pleasant vs. unpleasant). Simultaneously, an adversarial model is trained to predict sensitive or bias-related attributes (e.g., subcategories reflecting social concepts). The embeddings are updated to improve task performance while minimizing the adversary's ability to predict the bias attribute, effectively removing unwanted correlations. The optimization follows an alternating gradient update strategy. The task loss is computed using binary cross-entropy, while the adversarial loss uses standard cross-entropy to reduce bias attribute predictability. Training is conducted over 50 epochs (tunable), with a learning rate of 0.001 and a batch size of 64. This debiasing process is applied independently to text embeddings and image embeddings for creating the diagnostic baseline in our framework.

Metrics To evaluate model performance, we use **accuracy** as the primary metric, defined as the proportion of correct predictions over the total number of predictions.

For assessing fairness, we compute the **True Positive Rate (TPR)** across different groups defined by sensitive attributes (e.g., religion, nationality, disability, or sexual orientation). TPR measures the proportion of positive instances that are correctly identified by the model within each group. For a given group g, TPR is defined as:

$$\text{TPR}_g = \frac{\text{TP}_g}{\text{TP}_g + \text{FN}_g}$$

where TP_g and FN_g are the number of true positives and false negatives for group g, respectively. Disparities in TPR across groups are used to quantify representational or predictive bias.

For internal similarity analysis, we adopt the **Caliskan-style association test** (WEAT) (Caliskan et al., 2017b), a widely used metric for quantifying bias in embedding spaces. Given target sets X and Y (e.g., groups of concepts such as *Islam* and *Christianity*) and attribute sets A and B (e.g., *pleasant* and *unpleasant*), the WEAT score evaluates the differential association of X and Y with A and B using a similarity function $S(\cdot, \cdot)$ such as cosine similarity. The test statistic is defined as:

$$d(X, Y, A, B) = \sum_{x \in X} \left[\frac{1}{|A|} \sum_{a \in A} \mathcal{S}(x, a) - \frac{1}{|B|} \sum_{b \in B} \mathcal{S}(x, b) \right] - \sum_{y \in Y} \left[\frac{1}{|A|} \sum_{a \in A} \mathcal{S}(y, a) - \frac{1}{|B|} \sum_{b \in B} \mathcal{S}(y, b) \right] \quad (1)$$

This score reflects how much more strongly the target concepts in X are associated with attributes in A versus B, relative to the concepts in Y. A high absolute value of $d(X, Y, A, B)$ indicates greater bias. This test is applied independently to embeddings from each modality $M_i \in M$, using the similarity function $S(\cdot, \cdot)$ defined in our framework.

D Setting the Parameter ϵ

From a *representational similarity* perspective, we set $\epsilon = 0.1$ to define our thresholds for detecting emergent bias. The small threshold $\epsilon = 0.1$ ensures that the embeddings of different identity groups within each unimodal representation are highly similar, indicating minimal residual bias and effective unimodal debiasing. That is, the internal representations for different groups are nearly indistinguishable in the latent space, reflecting no meaningful separation based on sensitive attributes.

E Gender Bias: Objects Associated with Occupations

The VisoGender dataset (Hall et al., 2023) is a balanced dataset of image-caption pairs, composed of occupation-related images and captions produced via templates, which denote the gender and occupation of the subject. A subset of the dataset is composed of occupation-related images that contain a human subject (either male or female) and an object related to their occupation. Example objects include "stethoscope", associated with doctors, and "mixing spoon", associated with bakers. For a variety of objects in this subset, we calculated the average cosine similarity between male and female-associated samples for each object. In Figure 7, we illustrate the average cosine similarities for both text-only and multimodal embeddings, for four different objects: *computer*, *stock*, *notepad*, and *presentation*. Although we observe high similarity between text embeddings for all four objects, we observe low similarity between the multimodal embeddings for male vs. female subjects. In particular, we observe low similarity between multimodal embeddings for male and female subjects for *computer*, *notepad*, and *presentation*.

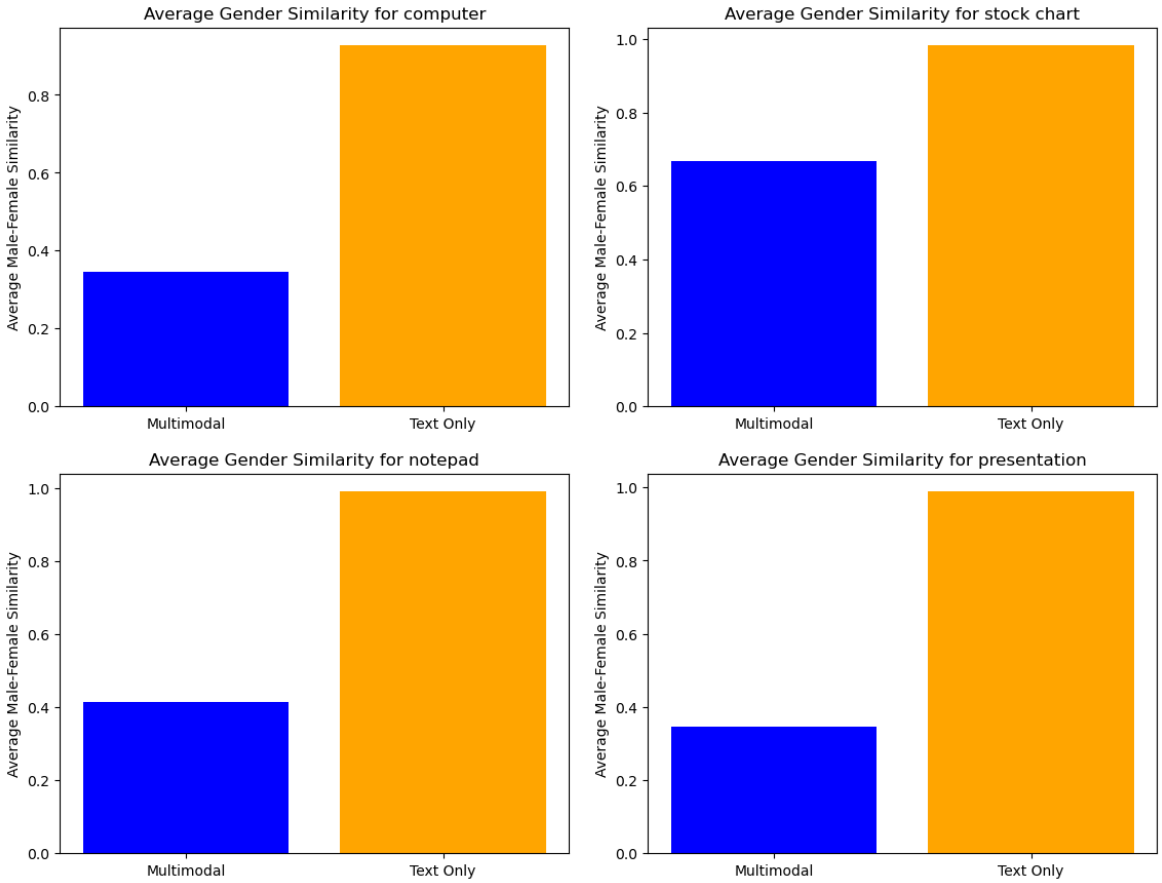

Figure 7: Cosine similarity between text-only and multimodal embeddings for occupation-related image/caption pairs containing female and male subjects. We observe high similarity in text-only embeddings for all objects, but low similarity between multimodal embeddings for image-caption pairs with male and female subjects.