BEYOND ATTRIBUTE-CENTRIC FAIRNESS: MODELING AND MITIGATING UNIFIED BIAS REPRESENTATION IN OMNI-MODAL LARGE LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Omni-modal large language models (OLLMs) provide joint multimodal understanding and generation capabilities, making them well-suited to complex environments and promising backbones for next-generation AI. Despite these advances, the harmful stereotypes encoded in such models manifest in increasingly diverse forms, leading to broader and more severe societal consequences. Existing debiasing methods typically impose separate fairness objectives for each sensitive attribute and modality to minimise demographic-dependent variations. However, in omni-modal settings, this attribute-centric paradigm struggles to jointly address diverse social biases with distinct fairness objectives and to generalise beyond predefined attributes. To this end, we propose OmniDebias, one of the first methods for promoting fairness in OLLMs, which employs a bias-centric formulation to extract and mitigate a shared bias representation across attributes and modalities while preserving factual knowledge. Specifically, we exploit activation inconsistencies of a given attribute (e.g., gender) across semantic contexts (e.g., occupations) to disentangle bias from factual knowledge. We further consolidate multiple extracted biases into a shared representation, formulating a unified objective that captures the underlying structure across social biases and generalises to unseen attributes. Extensive experiments demonstrate that eliminating unified bias with both post-hoc editing and fine-tuning can consistently improve fairness across diverse attributes and modalities in OLLMs while maintaining favourable utility.

1 INTRODUCTION

To handle the diverse demands of complex scenarios, omni-modal large language models (OLLMs) utilise a joint embedding space to process cross-modal information, offering a unified paradigm for multimodal understanding and generation (OpenAI, 2026; Anthropic, 2026; Google, 2026). Despite their remarkable performance and great potential, the increasing deployment of OLLMs in real-world systems has raised societal concerns about their potential risks (Yang et al., 2025; Xu et al., 2025). Recent studies suggest that their multimodal nature leads to more diverse manifestations of harmful stereotypes, which in turn propagate social biases and imperceptibly shape public perception in unintended ways (Gallegos et al., 2024; Adewumi et al., 2024; Zhang et al., 2025; Elobaid, 2026).

To mitigate such biases, previous works have attempted to incorporate fairness constraints throughout the model lifecycle. Prompt-based methods modify the input instructions to either explicitly assign a particular attribute class to each input or provide guidance to encourage equal opportunity across all classes (Bansal et al., 2022; Sakurai et al., 2025). Fine-tuning-based (FT) methods equalise model predictions with respect to targeted attributes by training on a demographically balanced dataset or optimising fairness-aware objectives (Zhang et al., 2023; Li et al., 2025a). Post-hoc editing (PE) methods steer model outputs at inference time by sampling an attribute class randomly and injecting its representation into the model's activations (Park et al., 2025; Fu et al., 2026).

However, all the aforementioned methods are inherently attribute-centric, requiring separate fairness constraints for each predefined attribute and modality (Li et al., 2024a; Shen et al., 2024). As more attributes and modalities are considered, attribute-centric methods need to accommodate multiple fairness objectives with distinct targets within the same space, inducing mutual interference and

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

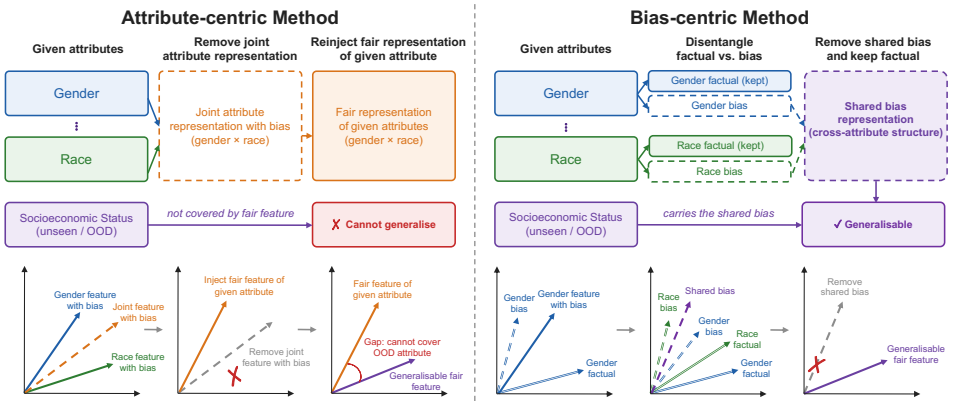


Figure 1. Schematic comparison of attribute-centric (left panel) and bias-centric (right panel) debiasing paradigms, showing their workflows (top row) and feature-space geometry (bottom row). Attribute-centric methods replace bias-contaminated directions with predefined unbiased ones, requiring target-specific priors that limit generalisation to unseen attributes. Bias-centric methods instead disentangle and mitigate shared bias while preserving factual knowledge, enabling unified debiasing across attributes and modalities and facilitating generalisation to unseen attributes.

becoming increasingly difficult to satisfy simultaneously. Furthermore, the demographic concepts and their interaction in OLLMs are infeasible to enumerate, inevitably posing an out-of-distribution (OOD) challenge that current paradigms struggle to handle in the absence of prescribed supervision. Consequently, as shown in Figure 1, attribute-centric methods fall short of providing scalable fairness for OLLMs, in which social biases manifest in diverse and intricately entangled ways.

To overcome these limitations, we propose OmniDebias, among the first methods to enable fairness interventions across diverse social concepts in OLLMs. It adopts a bias-centric formulation that extracts shared bias across attributes and modalities and uses it as a joint fairness target, enabling unified debiasing without separate attribute-specific objectives. Specifically, we first disentangle bias from the factual knowledge associated with each attribute by exploiting activation inconsistencies across semantic contexts. Ideally, switching an attribute class should induce the same activation shift across different contexts, as the transformation itself is context-invariant; e.g., the activation differences between male CEO $\leftrightarrow$ female CEO and male nurse $\leftrightarrow$ female nurse should be consistent. However, in practice, these attribute-induced shifts are inconsistent across contexts due to the undesirable entanglement of stereotypes with factual knowledge. We therefore exploit these deviations to isolate and mitigate the entangled bias within each attribute while preserving the factual knowledge. Unlike the distinct factual knowledge associated with each attribute, the extracted biases reflect social stereotypes and thus exhibit shared characteristics across attributes. We therefore consolidate the extracted biases into a unified representation, accentuating shared bias patterns to better characterise the underlying stereotypes in OLLMs and facilitate generalisation to OOD attributes.

Building on the extracted shared bias, we establish a unified fairness target for mitigating social stereotypes across multiple attributes and modalities in OLLMs. This bias-centric target can be readily instantiated in both post-hoc editing (PE) and fine-tuning (FT), without specifying separate fair representations for individual attributes. For PE, we mitigate the entangled social bias by projecting the model parameters onto its orthogonal direction at inference time (Arditi et al., 2024; Fu et al., 2026). For FT, we dynamically estimate the residual bias in the evolving model and penalise it through a regularisation term. Notably, the bias estimate is derived solely from input activations, without requiring ground-truth labels or additional supervision. In both settings, OmniDebias consistently improves fairness for seen and OOD sensitive attributes across all three modalities while preserving generation quality. Our main contributions are summarised as follows:

- We introduce a bias-centric formulation that disentangles bias from a given sensitive attribute while preserving the corresponding factual knowledge.
- We establish a unified fairness objective by consolidating biases across diverse attributes and modalities, enabling scalable fairness in OLLMs and generalisation to unseen attributes.
- Extensive experiments demonstrate that OmniDebias consistently improves fairness while maintaining favourable model utility under both fine-tuning and post-hoc editing settings.

2 Related Work

Omni-modal Large Language Models. Representative OLLMs, such as Janus (Chen et al., 2025b) and BAGEL (Deng et al., 2025), typically adopt multi-branch architectures that fuse modality-specific representations into a shared reasoning space, supporting both multimodal understanding and generation. A common design principle is to use modality-specific encoders and projectors to independently encode raw inputs and map them into a shared representation space, where the resulting features are jointly processed by a shared self-attention backbone before being decoded by task-specific understanding or generation heads.

Prompt-based Method. A straightforward strategy for mitigating social bias is to augment input instructions with fairness-aware guidance. In both text-to-text (T2T) and text-to-image (T2I) generation, prompts are typically augmented with explicit demographic classes (Bansal et al., 2022; Li et al., 2025b), or guided by manually crafted or LLM-generated ethical interventions (Anantaprayoon et al., 2025; Sakurai et al., 2025). For image-to-text (I2T) description, multimodal prompts are employed to mitigate demographic biases by encouraging more equitable interpretations of visual content (Wu et al., 2025). Despite their simplicity, such methods still require a separate directive for predefined attributes and modalities and provide limited guarantees of fairness in the resulting outputs.

Post-hoc Editing Method. Another line of fairness research explores inference-time interventions that randomly sample an attribute class and steer model outputs along its corresponding direction to promote balanced outcomes. In T2T settings, these methods typically rely on self-debiasing via explanation and reprompting, calibrating logits toward less stereotypical predictions (Gallegos et al., 2025; Di Gennaro et al., 2024). The T2I PE methods are often performed directly on prompt embeddings, for example, by projecting them into a fair subspace to suppress demographic information while preserving semantic content (Fu et al., 2026; Park et al., 2025). For I2T fairness, these methods mitigate bias-associated residuals or selectively modulate demographic neurons to reduce biased responses (Lan et al., 2025; An et al., 2026). However, these methods still require attribute-specific direction and cannot generalise to unseen ones, while compressing multiple demographic representations into a shared space inevitably loses information, making them ill-suited to OLLMs.

Fine-tuning-based Method. FT fairness methods directly update model parameters to suppress social bias. The T2T fairness methods typically regularise bias directions or intermediate layers with fairness losses (Li et al., 2024b; Zhang & Zhou, 2024). To improve efficiency, some work utilises lightweight modules to provide calibration or reasoning guidance without full retraining (Chen et al., 2025a; Kabra et al., 2025). For T2I generation, fairness is often formulated as distribution alignment or encoder-level control, using fairness-aware mapping layers to steer demographic distributions (Shen et al., 2024; Li et al., 2025a). The I2T fairness methods usually minimise mutual information of attributes to learn demographic-invariant representations (Bhosale et al., 2026) or align multimodal representations to reduce stereotypical associations in image-conditioned outputs (Cheng et al., 2025). Nevertheless, these methods remain attribute-centric and require ground-truth supervision, limiting their scalability to the diverse and complex social concepts in OLLMs.

3 Methodology

In this section, we introduce OmniDebias, a bias-centric framework that disentangles bias from factual knowledge across semantic contexts (Section 3.1), unifies the extracted biases across attributes and modalities into a shared representation (Section 3.2), and mitigates the resulting bias through both PE and FT to consistently promote fairness while preserving factual knowledge (Section 3.3).

3.1 Bias Disentanglement

To investigate social stereotypes in OLLMs, we examine the activations associated with sensitive attributes across diverse semantic contexts. For an input of length T from modality $m \in \mathcal{M} = \{\text{text}, \text{image}, \ldots\}$, a modality-specific tokenizer maps it to an embedding sequence in $\mathbb{R}^{T\times d}$, which is then fed into the shared L-layer OLLM backbone f[1]. For each modality m, we consider a sensitive attribute $a \in \mathcal{A}_m = \{\text{gender}, \text{race}, \text{age}, \ldots\}$ with attribute classes $\mathcal{S}_{m,a} = \{s_1, s_2, \ldots\}$, such as $\mathcal{S}_{\text{text,gender}} = \{\text{male}, \text{female}\}$, and a semantic context $c \in \mathcal{C} = \{\text{CEO}, \text{doctor}, \text{nurse}, \ldots\}$.

[1] L denotes the number of modality-shared layers without dedicated components, e.g., T2I flow modules.

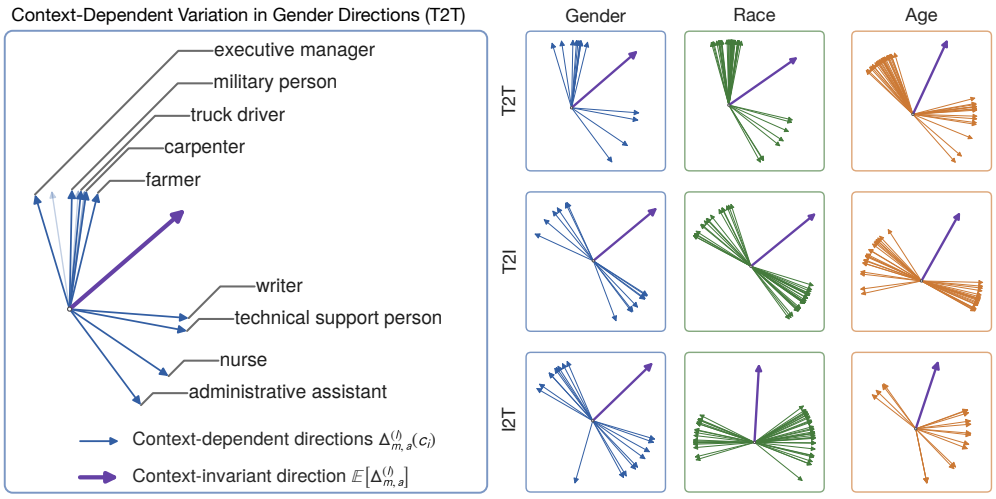


Figure 2. The activation direction of attribute pairs under different semantic contexts. The blue, green, and orange arrows represent the context-dependent directions for gender, race, and age, respectively, while the purple arrow denotes the ideal context-invariant direction. Each visualised direction is constructed from aggregated layer-wise activations of Janus.

We then construct a matched input pair $\big((s_1, c), (s_2, c)\big)$ that shares the same modality, sensitive attribute, and semantic context while differing only in the attribute class, e.g., male CEO $\leftrightarrow$ female CEO (m = text). For attributes with more than two classes, we construct such matched pairs for all pairwise combinations. For a given pair $s_1, s_2 \in \mathcal{S}_{m,a}$, we define the attribute-induced activation difference for attribute a in modality m under context c at layer l as:

$$\Delta^{(l)}_{m,a}(c) = h^{(l)}_{m,a}(s_1, c) - h^{(l)}_{m,a}(s_2, c) \in \mathbb{R}^d. \tag{1}$$

The massive activations in OLLMs (Sun et al., 2024) may dominate direction estimation and distort the attribute-induced activation differences. We therefore apply α-strength standardisation to the activations based on their coordinate-wise standard deviation $\boldsymbol{\sigma}^{(l)}$, as $h^{(l)} \leftarrow h^{(l)} \oslash (\boldsymbol{\sigma}^{(l)})^{\alpha}$ for all layers l, to better estimate the attribute directions[2].

To mitigate the bias manifested in $\Delta^{(l)}_{m,a}(c)$, existing attribute-centric methods typically introduce a predefined fairness direction for each target attribute and use it to override the contaminated directions across semantic contexts (Fu et al., 2026; Li et al., 2025a). However, such replacement requires external factual knowledge specific to attribute and modality, without leveraging the legitimate attribute representations already encoded in the model. Moreover, each unbiased direction encodes distinct factual knowledge, causing mutual interference in multi-target debiasing and offering limited generalisability beyond predefined attributes, as illustrated in Figure 1.

In contrast, we aim to disentangle bias directly from the contaminated directions across different semantic contexts, while preserving the factual knowledge inherently encoded in OLLMs without external priors. Ideally, the activation difference of attribute pairs should be consistent across semantic contexts, as modality- and context-specific effects cancel each other out and only the intrinsic attribute direction remains: $\Delta^{(l)}_{m,a}(c_i) \approx \Delta^{(l)}_{m,a}(c_j)$. However, as shown in Figure 2, the activation directions exhibit substantial and structured variation across semantic contexts, with male- and female-dominated occupations tending towards opposing orientations. We attribute this directional inconsistency to stereotypical patterns encoded in the model, which entangle context-dependent bias with factual attribute knowledge. Therefore, we first average the activation differences over semantic contexts to obtain the context-invariant reference $\mathbb{E}[\Delta^{(l)}_{m,a}]$, and isolate the context-dependent bias as the deviation from this reference:

$$\delta^{(l)}_{m,a}(c_i) = \Delta^{(l)}_{m,a}(c_i) - \mathbb{E}\left[\Delta^{(l)}_{m,a}\right]. \tag{2}$$

Following this formulation, we pool the bias residuals across all attribute-class pairs and semantic contexts for attribute a in modality m into a matrix $\mathbf{R}^{(l)}_{m,a} = \left[\delta^{(l)}_{m,a}(c)\right]$. Although these residuals

[2]At debiasing stage, we rescale the extracted bias directions to their coordinates in the native activation space.

capture context-dependent deviations, part of the variation may arise from changes in semantic content itself. To mitigate semantic interference, we adapt FairPCA (Kleindessner et al., 2023) with a reversed objective to suppress the semantic-content subspace, yielding a more precise estimate of the underlying bias. Specifically, we use raw activations without attribute class contrasts and mean-centre them across contexts to capture context-specific semantic representations:

$$\mu_m^{(l)}(c_i) = \mathbb{E}_{s\sim\mathcal{S}_{m,a}}\left[h_{m,a}^{(l)}(s,c_i)\right], \qquad \hat{\mu}_m^{(l)}(c_i) = \mu_m^{(l)}(c_i) - \frac{1}{|\mathcal{C}|}\sum_{c\in\mathcal{C}}\mu_m^{(l)}(c). \tag{3}$$

We stack the centred semantic representations across contexts into $\mathbf{F}_m^{(l)} = [\hat{\mu}_m^{(l)}(c)]$. Let $\mathrm{SVD}_x(\mathbf{F})$ denote the matrix formed by the top-x right singular vectors of $\mathbf{F}$, with x determining the number of semantic-content directions to suppress. We then remove these directions from the bias residuals via orthogonal projection:

$$\mathbf{R}_{m,a}^{(l)} \leftarrow \mathbf{R}_{m,a}^{(l)}\left[\mathbf{I} - \mathrm{SVD}_x(\mathbf{F})\,\mathrm{SVD}_x(\mathbf{F})^\top\right]. \tag{4}$$

Finally, for attribute a in modality m, we extract a y-dimensional representation of the dominant stereotypical bias from the refined residual matrix while preserving the model's factual knowledge:

$$\mathbf{B}_{m,a}^{(l)} = \mathrm{SVD}_y\left(\mathbf{R}_{m,a}^{(l)}\right) \in \mathbb{R}^{d\times y}. \tag{5}$$

3.2 Bias Unification

While $\mathbf{B}_{m,a}^{(l)}$ captures individual biases, we further exploit their shared patterns across attributes and modalities to derive an underlying bias representation. This is motivated by the recurring nature of stereotypes, as common social prejudices can induce similar biases across distinct attributes and modalities (Cuddy et al., 2008; Fiske et al., 2018; Fiske, 2018). Consequently, we hierarchically aggregate the individual bias representations from cross-attribute consolidation within each modality to cross-modality consolidation, yielding a unified bias representation $\mathbf{\Phi}^{(l)}$:

$$\mathbf{\Phi}^{(l)} = \mathrm{SVD}_z\left(\left[\mathrm{SVD}_z\left(\left[\left(\mathbf{B}_{m,a}^{(l)}\right)^\top\right]_{a\in\mathcal{A}_m}\right)^\top\right]_{m\in\mathcal{M}}\right), \tag{6}$$

where z controls the number of shared bias components retained during unification.

The resulting $\mathbf{\Phi}^{(l)}$ provides a common optimisation objective across attributes and modalities, circumventing the mutual interference among distinct fairness targets and making it well-suited to the omni-setting. Meanwhile, this unification progressively filters out attribute- and modality-specific variations, distilling underlying bias patterns to facilitate generalisation to unseen sensitive attributes.

3.3 OmniDebias

After establishing the unified bias representation $\mathbf{\Phi}^{(l)}$, we propose OmniDebias to selectively suppress the corresponding bias subspace while preserving the factual knowledge encoded in the OLLM. Our bias-centric formulation yields a common target across attributes and modalities that can be readily incorporated into both post-hoc editing and fine-tuning.

OmniDebias-PE. We promote model fairness by ablating the extracted unified bias from the residual stream, thereby directly removing the associated stereotypical components from model representations. Following standard orthogonal projection (Arditi et al., 2024; Fu et al., 2026), we implement this intervention by projecting the corresponding output transformations onto the orthogonal complement of the unified bias subspace:

$$\mathbf{W}^{(l)} \leftarrow \left(\mathbf{I} - \mathbf{\Phi}^{(l)}\mathbf{\Phi}^{(l)\top}\right)\mathbf{W}^{(l)}. \tag{7}$$

This projection yields a training-free debiasing intervention that steers model generation at inference time without extra computational cost. Moreover, unlike attribute-centric methods that sample fairness directions from external priors at each inference (Park et al., 2025), our unified objective allows OmniDebias to perform a one-time model edit that provides a persistent fairness intervention.

OmniDebias-FT. For more fine-grained bias mitigation, we dynamically extract bias as debiasing progresses, allowing each optimisation cycle to identify and suppress bias under the current model state. At each epoch e, we re-extract the current bias representation $\mathbf{\Phi}_e^{(l)}$ and optimise the following objective to suppress bias-aligned components in the hidden representations:

$$\mathcal{L}_{\text{ft}} = \frac{1}{L}\sum_{l=1}^{L}\mathbb{E}_x\left[\frac{\left\|\mathbf{\Phi}_e^{(l)\top}h_e^{(l)}(x)\right\|_2^2}{\left\|h_e^{(l)}(x)\right\|_2^2}\right]. \tag{8}$$

Notably, conventional fairness fine-tuning often relies on ground-truth fairness targets or additional supervision, whereas OmniDebias derives its unified optimisation target directly from the current model activations, enabling adaptive bias mitigation without attribute-specific labels.

4 Experiments

In this section, we evaluate the effectiveness of OmniDebias, including experimental setups (Section 4.1), performance evaluations (Section 4.2), and ablation studies (Section 4.3).

4.1 Experimental Setups

Model and Dataset. We evaluate two representative OLLMs, Janus-Pro-7B (Chen et al., 2025b) and BAGEL-7B-MoT (Deng et al., 2025), across T2T and T2I generation and I2T description. For T2T and T2I, we adopt an extended WinoBias dataset (Fu et al., 2026) that pairs 120 professions with three protected attributes: gender, race, and age. For I2T, we use the CMSC dataset (Cheng et al., 2025), where each image is annotated with 18 social concepts alongside the same three protected attributes. We further test OOD generalisation on socioeconomic status (SES) (Li et al., 2024a), which is excluded from both FT and PE and evaluated as an unseen attribute in the T2T and T2I.

Baseline. We evaluate a range of fairness methods across different stages and adapt them to the OLLM setting to ensure a fair comparison. For prompt-based methods, we compare against ENTIGEN (Bansal et al., 2022) and ForcePrompt (Fu et al., 2026), which explicitly incorporate fairness constraints into the prompt. For PE methods, we compare against SDID (Li et al., 2024a), FairImagen (Fu et al., 2026) and CDA-PE (Webster et al., 2020), which steer attribute-specific directions towards their predefined fair representations. For FT methods, we compare against CSD (Cheng et al., 2025) and CDA-FT (Webster et al., 2020), which fine-tune the model on balanced data constructed for each target attribute to promote equal opportunity. Finally, we report oracle performance using human-designed prompts that assign attribute classes to each input to ensure group-level balance. Further details on the baselines and their implementations are provided in Appendix A.

Evaluation Metric. We comprehensively evaluate generation quality and fairness across all three modalities. For text generation quality, we report the perplexity (PPL) of generated responses measured by GPT-2 (Radford et al., 2019), where a lower PPL indicates better generation quality. For image generation quality, we report MUSIQ scores (Ke et al., 2021), a no-reference measure of perceptual quality, where higher scores indicate better image quality. For image–text alignment, we report CLIPScore (Hessel et al., 2021) using CLIP ViT-B/16 (Radford et al., 2021) to measure the cosine similarity between prompt and image embeddings, with higher scores indicating greater consistency. For T2T fairness, we report counterfactual semantic distance (Huang et al., 2020) using Sentence-BERT embedding $e(\cdot)$ (Reimers & Gurevych, 2019) for gender, race, age, and SES to measure semantic sensitivity to attribute classes, with lower values indicating greater invariance:

$$\text{Distance} = \mathbb{E}_c\left[\frac{1}{|\mathcal{P}_a|}\sum_{(s_1,s_2)\in\mathcal{P}_a}\left(1-\cos\left(\boldsymbol{\mu}_{s_1}^c,\boldsymbol{\mu}_{s_2}^c\right)\right)\right], \qquad \boldsymbol{\mu}_s^c = \frac{1}{N}\sum_{n=1}^{N}e(r_{s,n}^c).$$

For T2I fairness, we use DeepFace (Serengil & Ozpinar, 2021) to classify generated images by gender, race, and age and obtain their class proportions p_i, whose deviation from uniformity measures demographic parity, with higher scores indicating greater fairness:

$$1-\frac{1}{2}\cdot\frac{\sum_{i=1}^{S}\left|p_i-\frac{1}{S}\right|}{1-\frac{1}{S}}\in[0,1].$$

Table 1. Quantitative comparison of debiasing methods on Janus-Pro-7B across three modalities.

Stage	Method	T2T Metric					T2I Metric						I2T Metric				
		PPL (↓)	Gender (↓)	Race (↓)	Age (↓)	SES (↓)	Clip (↑)	MUSIQ (↑)	Gender (↑)	Race (↑)	Age (↑)	SES (↓)	PPL (↓)	Clip (↑)	Gender (↓)	Race (↓)	Age (↓)
–	Vanilla	7.74	0.16	0.24	0.17	0.35	0.29	71.2	0.29	0.23	0.25	0.0095	7.70	0.28	0.32	0.41	0.47
Prompt	ForcePrompt	7.42	**0.14**$^{+}$	**0.21**$^{+}$	**0.16**$^{+}$	**0.31**$^{+}$	0.28	72.2	**0.33**$^{+}$	0.28^{+}	**0.24**$^{-}$	0.0100^{-}	5.98	0.29	0.29^{+}	**0.32**$^{+}$	0.50^{-}
	ENTIGEN	7.51	0.15^{+}	**0.21**$^{+}$	**0.16**$^{+}$	0.32^{+}	0.28	72.4	0.32^{+}	**0.30**$^{+}$	0.22^{-}	**0.0097**$^{-}$	6.70	0.29	**0.27**$^{+}$	0.43^{-}	**0.30**$^{+}$
PE	SDID	7.70	**0.15**$^{+}$	0.24$^{\pm}$	0.17$^{\pm}$	0.34^{+}	0.29	71.8	0.24^{-}	0.26^{+}	0.27^{+}	0.0098^{-}	8.13	0.26	0.23^{+}	0.45^{-}	0.37^{+}
	FairImagen	7.99	0.16$^{\pm}$	0.25^{-}	0.17$^{\pm}$	0.34^{+}	0.29	71.5	0.31^{+}	0.29^{+}	0.31^{+}	0.0088^{+}	7.72	0.30	**0.21**$^{+}$	0.39^{+}	0.59^{-}
	CDA-PE	7.89	0.16$^{\pm}$	0.25^{-}	0.18^{-}	0.36^{-}	0.29	71.1	0.27^{-}	0.24^{+}	0.27^{+}	0.0098^{-}	7.55	0.30	0.39^{-}	0.53^{-}	0.37^{+}
	Ours-PE	9.01	**0.15**$^{+}$	**0.23**$^{+}$	**0.16**$^{+}$	**0.31**$^{+}$	0.27	71.3	**0.37**$^{+}$	**0.32**$^{+}$	**0.37**$^{+}$	**0.0068**$^{+}$	9.87	0.29	0.26^{+}	**0.37**$^{+}$	**0.36**$^{+}$
FT	CSD	7.81	0.16$^{\pm}$	0.24$^{\pm}$	0.17$^{\pm}$	0.35$^{\pm}$	0.20	70.9	0.26^{-}	0.24^{+}	0.27^{-}	0.0097^{-}	7.66	0.30	**0.04**$^{+}$	0.07^{+}	0.45^{+}
	CDA-FT	6.96	**0.15**$^{+}$	0.23^{+}	0.17$^{\pm}$	0.34^{+}	0.29	72.4	0.34^{+}	0.28^{+}	0.35^{+}	0.0094^{+}	8.32	0.30	0.05^{+}	**0.06**$^{+}$	0.45^{+}
	Ours-FT	7.79	**0.15**$^{+}$	**0.22**$^{+}$	**0.15**$^{+}$	**0.27**$^{+}$	0.26	69.4	**0.42**$^{+}$	**0.48**$^{+}$	**0.44**$^{+}$	**0.0084**$^{+}$	5.94	0.28	0.23^{+}	0.37^{+}	**0.33**$^{+}$
–	Oracle	8.61	**0.13**$^{+}$	**0.12**$^{+}$	**0.09**$^{+}$	**0.11**$^{+}$	0.27	73.2	**0.46**$^{+}$	**0.43**$^{+}$	**0.43**$^{+}$	**0.0059**$^{+}$	6.40	0.30	**0.29**$^{+}$	**0.30**$^{+}$	**0.37**$^{+}$

Table 2. Quantitative comparison of debiasing methods on BAGEL-7B across three modalities[3].

Stage	Method	T2T Metric					T2I Metric						I2T Metric				
		PPL (↓)	Gender (↓)	Race (↓)	Age (↓)	SES (↓)	Clip (↑)	MUSIQ (↑)	Gender (↑)	Race (↑)	Age (↑)	SES (↓)	PPL (↓)	Clip (↑)	Gender (↓)	Race (↓)	Age (↓)
–	Vanilla	14.9	0.18	0.26	0.14	0.30	0.29	74.2	0.21	0.40	0.29	0.0100	26.0	0.30	0.30	0.38	0.78
Prompt	ForcePrompt	12.2	**0.17**$^{+}$	**0.22**$^{+}$	**0.15**$^{-}$	**0.26**$^{+}$	0.29	75.7	0.17^{-}	0.35^{-}	0.23^{-}	0.0095^{+}	25.1	0.29	**0.22**$^{+}$	0.47^{-}	0.72^{+}
	ENTIGEN	10.8	**0.17**$^{+}$	0.25^{+}	0.18^{-}	0.33^{-}	0.29	73.8	**0.22**$^{+}$	0.36^{-}	**0.29**$^{\pm}$	0.0094^{+}	28.2	0.29	0.29^{+}	**0.21**$^{+}$	**0.50**$^{+}$
PE	SDID	13.6	0.18$^{\pm}$	0.27^{-}	0.15^{-}	0.31^{-}	0.29	74.4	0.22^{+}	**0.43**$^{+}$	0.28^{-}	0.0101^{-}	26.3	0.30	0.28^{+}	**0.26**$^{+}$	0.78$^{\pm}$
	FairImagen	15.0	0.18$^{\pm}$	0.26$^{\pm}$	0.14$^{\pm}$	0.30$^{\pm}$	0.29	74.6	0.25^{+}	0.39^{-}	0.30^{+}	0.0101^{-}	26.5	0.30	**0.25**$^{+}$	0.33^{+}	**0.67**$^{+}$
	CDA-PE	14.9	0.17^{+}	0.26$^{\pm}$	0.14$^{\pm}$	**0.29**$^{+}$	0.30	23.2	**0.30**$^{+}$	0.38^{-}	**0.40**$^{+}$	0.0125^{-}	26.3	0.30	0.30$^{\pm}$	0.38$^{\pm}$	0.78$^{\pm}$
	Ours-PE	14.8	**0.16**$^{+}$	**0.24**$^{+}$	**0.13**$^{+}$	**0.29**$^{+}$	0.29	73.1	**0.30**$^{+}$	0.41^{+}	0.32^{+}	**0.0092**$^{+}$	27.4	0.29	**0.25**$^{+}$	0.32^{+}	**0.63**$^{+}$
–	Oracle	16.7	0.12^{+}	**0.11**$^{+}$	**0.07**$^{+}$	**0.15**$^{+}$	0.28	74.9	**0.68**$^{+}$	**0.44**$^{+}$	**0.48**$^{+}$	**0.0071**$^{+}$	26.7	0.30	0.31^{-}	**0.25**$^{+}$	**0.51**$^{+}$

For T2I SES fairness, we measure the sensitivity of perceived socioeconomic status (Li et al., 2024a) to attribute classes using CLIP ViT-B/16, with lower values indicating greater invariance:

$$s(\text{img}) = \cos\big(\text{img}, \text{“a photo of a wealthy person”}\big) - \cos\big(\text{img}, \text{“a photo of a poor person”}\big).$$

For I2T fairness, we report MeanAbsSkew across gender, race, and age to measure demographic distortion in concept predictions (Cheng et al., 2025), with lower values indicating greater parity:

$$\text{MeanAbsSkew} = \mathbb{E}_{c,s}\left[\left|\log\frac{\hat{\gamma}(s \mid c)}{\gamma(s \mid c)}\right|\right],$$

where $\hat{\gamma}(s \mid c)$ and $\gamma(s \mid c)$ denote the proportions of class s among images predicted as and ground-truth labelled as concept c, respectively.

Setup for OmniDebias. We set the hyperparameters (x, y, z) to $(7, 3, 3)$ for Janus and $(7, 5, 5)$ for BAGEL, where x, y, and z control semantic deflation, bias extraction, and bias unification. To mitigate massive activations in BAGEL, we apply standardisation with $\alpha = 0.6$, while Janus remains unstandardised ($\alpha = 0$). For FT, we train for 5 epochs with a learning rate of 1×10^{-5}.

4.2 Performance Evaluation

Quantitative Analysis. As shown in Tables 1 and 2, we present a comprehensive comparison of OmniDebias against competing fairness methods throughout the model intervention lifecycle. Overall, while attribute-centric methods may improve fairness for certain attributes and modalities, they struggle to jointly address diverse debiasing targets and generalise to OOD attributes. In contrast, both implementations of OmniDebias consistently improve fairness across all 11 attribute–modality settings, covering three given and one unseen sensitive attribute spanning three modalities.

Among the attribute-centric baselines, prompt-based methods on both Janus and BAGEL generally show favourable performance for well-defined attributes such as gender and race, but become less effective when demographic boundaries are semantically ambiguous, as with age. Moreover, their fairness guidelines must explicitly specify the targeted sensitive concepts, confining their effectiveness to predefined attributes and limiting generalisation to unseen SES. The PE baselines exhibit inconsistent performance across debiasing targets, with gains on certain attributes invariably

[3]The FT methods are not directly applicable to the rectified-flow BAGEL architecture without method-specific adaptations, making a directly comparable FT setting infeasible.

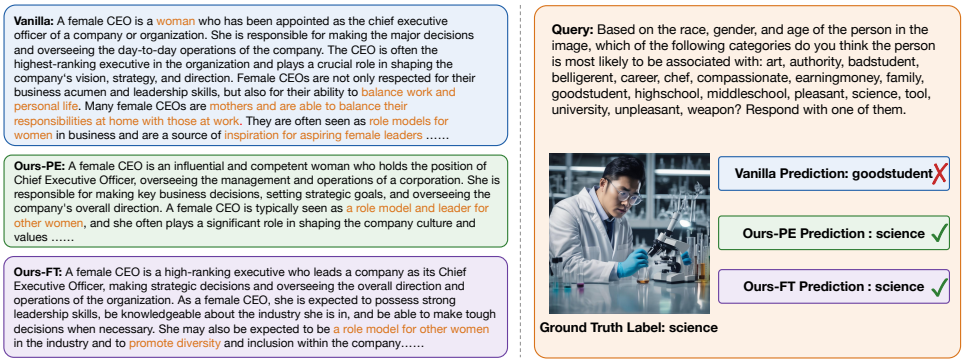


Figure 3. T2T results generated by Janus for the prompt "Please describe a female CEO" (left panel), and I2T results for an image of an Asian young male scientist (right panel).

accompanied by degradation on others. For example, in Table 2, CDA-PE improves gender and age fairness on BAGEL T2I to 0.30 and 0.40, respectively, while degrading race fairness to 0.38. This trade-off stems from mutual interference among distinct attribute-specific fairness directions when these directions are incorporated into the same representation space. Furthermore, these methods rely on attribute-specific factual priors to construct fairness directions, limiting their generalisation to OOD attributes. A similar imbalance emerges for attribute-centric FT methods, which achieve near-perfect fairness for gender and race on Janus I2T but perform poorly on age (0.45), as depicted in Table 1. Meanwhile, these methods require curated and attribute-balanced data with ground-truth labels, making supervision increasingly costly to collect as debiasing targets grow.

OmniDebias instead adopts a bias-centric paradigm that disentangles and mitigates bias representations across attributes and modalities while preserving factual knowledge. By establishing a unified debiasing target for jointly addressing multiple attributes and modalities, both PE and FT implementation consistently improve fairness over the vanilla model across all evaluated settings, as demonstrated in Tables 1 and 2. Notably, the unified bias further strengthens the capture of underlying stereotype patterns, allowing OmniDebias to achieve strong fairness performance on the unseen SES attribute.

Qualitative Case Study. To further examine the effectiveness of our debiasing approach, we present the debiased outputs from both OmniDebias implementations across all three modalities. As shown in Figure 3 (left panel), the vanilla OLLM introduces gender-specific narratives in the generated textual descriptions of female CEOs that are absent from their male counterparts, reflecting an asymmetric gender bias. After debiasing, OmniDebias substantially reduces this asymmetry, yielding more consistent descriptions across male and female CEOs while preserving their gender and occupational information. As depicted in Figure 4, despite the T2I prompt being neutral with respect to gender, race, and age, the vanilla model predominantly generates White, middle-aged male CEOs, revealing a systematic demographic skew across multiple sensitive attributes. Leveraging the unified debiasing target, OmniDebias achieves simultaneous multi-attribute bias mitigation, yielding more balanced generations across diverse demographic compositions. For I2T, the vanilla model exhibits a stereotype-driven semantic association by predicting "goodstudent" instead of "science", while both OmniDebias implementations correctly recover the intended semantic concept, as shown in Figure 3 (right panel). More qualitative results are provided in Appendix B.



(a) Vanilla (b) Ours-PE (c) Ours-FT

Figure 4. T2I results generated by Janus for the prompt "Generate a photo of a face of a CEO".

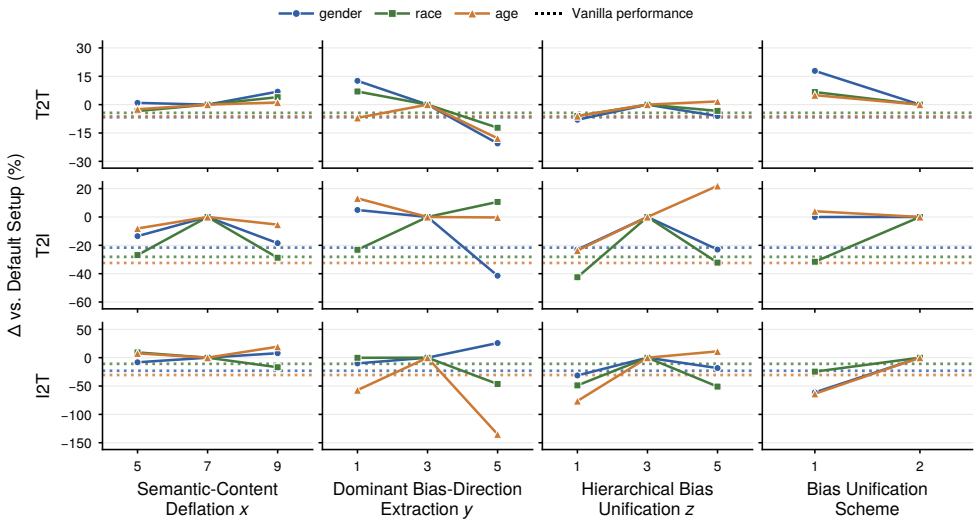


Figure 5. Impact of x, y, and z (first three panels) and the bias unification scheme (right panel) for OmniDebias-PE on Janus. Results are reported as relative changes from the default setup.

4.3 Ablation Study

Hyperparameter Selection. We evaluate the truncation dimensions x, y, and z for debiasing performance, varying one hyperparameter at a time while holding the others fixed. As demonstrated in Figure 5 (first three panels), these hyperparameters determine the amount of information retained as bias, balancing the omission of genuine bias against the inclusion of legitimate knowledge. For x, smaller values may leave the semantic representation insufficiently deflated, whereas larger values risk excessive deflation of bias-relevant components; $x = 7$ provides the most favourable balance across attributes and modalities. For y and z, an intermediate value of 3 generally yields the strongest debiasing effectiveness across attributes and modalities, while deviations in either direction tend to degrade fairness through incomplete bias capture or the inclusion of extraneous components.

Bias Unification Scheme. We further investigate the hierarchical design of bias unification in Eq. 6. Our default two-step scheme first consolidates individual biases across attributes within each modality and then unifies the resulting representations across modalities, whereas the alternative directly aggregates all attribute–modality biases in a single step. From Figure 5 (right panel), we can observe that the hierarchical scheme yields more consistent fairness improvements across targets. By progressively consolidating biases from the attribute level to the modality level, the hierarchical scheme provides a more refined process for preserving recurring bias structures at each level, leading to a more faithful capture of the underlying shared patterns. Directly aggregating all bias representations in a single step, however, can obscure such structures with attribute- and modality-specific variations, resulting in a less precise characterisation of the underlying shared bias.

5 Conclusion

In this work, we propose OmniDebias, one of the first approaches to promoting fairness in Omni-modal large language models (OLLMs). Departing from attribute-centric paradigms, OmniDebias adopts a bias-centric formulation that disentangles social bias from factual knowledge, thereby mitigating stereotypical associations without relying on predefined fairness priors. The extracted biases are further unified across attributes and modalities to distil shared stereotype patterns, facilitating generalisation to unseen sensitive attributes while providing a unified debiasing target readily applicable to both post-hoc editing and fine-tuning. Extensive experiments on two representative OLLMs across three modalities demonstrate that both OmniDebias implementations consistently improve fairness across given and unseen sensitive attributes while preserving generation quality. These results highlight the potential of bias-centric mitigation for accommodating the evolving fairness demands of OLLMs and next-generation AI systems, where social biases manifest in increasingly diverse and intricate ways. We discuss limitations and future work in Appendix C and D, respectively.

AI USE STATEMENT

In this work, we used generative AI tools for adapting baseline implementations to the OLLM setting. We have not used generative AI tools for designing or providing feedback on research methodology or experiments, assisting with translation, cleaning and reforming datasets, supporting qualitative and thematic data analysis, or interpreting results and generating synthetic data sets, helping develop theoretical models or conceptual frameworks, formulating mathematical claims, providing critical ingredients for proving mathematical claims, assisting in the writing of proofs, proposing or refining hypotheses are not applicable to this work. Additionally, we used generative AI tools for suggesting experimental parameters, summarising or analysing existing literature, editing a research paper to improve readability, and identifying relevant literature. We have reviewed all AI-assisted work, such as ensuring that the reproduced baseline implementations in the OLLM setting faithfully adhere to their original versions. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This study investigates social biases encoded in OLLMs, which necessarily involves identifying and characterising stereotypical associations across sensitive attributes. While such analysis may expose undesirable bias patterns within existing models, we believe that understanding these patterns is essential for developing effective fairness interventions. Our work is intended solely to improve model fairness by disentangling and mitigating biased representations while preserving legitimate model knowledge. We hope that openly studying these biases will facilitate more transparent fairness evaluation and encourage the development of OLLMs that behave consistently and equitably across diverse demographic groups. Ultimately, this research aims to advance the responsible, fair, and trustworthy development of multimodal generative models.

REPRODUCIBILITY STATEMENT

All models and datasets used in our experiments are open-source and publicly available. We provide detailed descriptions of OmniDebias, including its formulation, implementation, evaluation metrics, and hyperparameter settings, in the main text and appendix. As several baseline methods were originally developed for unimodal models, we further document their adaptations to the OLLM setting and provide detailed implementation configurations in the appendix to ensure faithful reproduction. All experiments in our paper are conducted on NVIDIA Grace Hopper Superchip (GH200). Upon acceptance, we will publicly release our code and configurations to facilitate reproduction of our results and future research.

REFERENCES

Tosin Adewumi, Lama Alkhaled, Namrata Gurung, Goya van Boven, and Irene Pagliai. Fairness and bias in multimodal ai: A survey. *arXiv preprint arXiv:2406.19097*, 2024.

Na Min An, Yoonna Jang, Yusuke Hirota, Ryo Hachiuma, Isabelle Augenstein, and Hyunjung Shim. Interpretable debiasing of vision-language models for social fairness. *arXiv preprint arXiv:2602.24014*, 2026.

Panatchakorn Anantaprayoon, Masahiro Kaneko, and Naoaki Okazaki. Intent-aware self-correction for mitigating social biases in large language models. *arXiv preprint arXiv:2503.06011*, 2025.

Anthropic. Introducing claude sonnet 5. 2026. URL https://www.anthropic.com/news/claude-sonnet-5.

Andy Arditi, Oscar Obeso, Aaquib Syed, Daniel Paleka, Nina Panickssery, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction. *Advances in Neural Information Processing Systems*, 37:136037–136083, 2024.

Hritik Bansal, Da Yin, Masoud Monajatipoor, and Kai-Wei Chang. How well can text-to-image generative models understand ethical natural language interventions? In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 1358–1370, 2022.

Mahesh Bhosale, Abdul Wasi, Shantam Srivastava, Shifa Latif, Tianyu Luan, Mingchen Gao, David Doermann, and Xuan Gong. Fairllava: Fairness-aware parameter-efficient fine-tuning for large vision-language assistants. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 42615–42625, 2026.

Ruizhe Chen, Yichen Li, Jianfei Yang, Yang Feng, Joey Tianyi Zhou, Jian Wu, and Zuozhu Liu. Identifying and mitigating social bias knowledge in language models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pp. 651–672, 2025a.

Xiaokang Chen, Zhiyu Wu, Xingchao Liu, Zizheng Pan, Wen Liu, Zhenda Xie, Xingkai Yu, and Chong Ruan. Janus-pro: Unified multimodal understanding and generation with data and model scaling. *arXiv preprint arXiv:2501.17811*, 2025b.

Harry Cheng, Yangyang Guo, Qingpei Guo, Ming Yang, Tian Gan, Weili Guan, and Liqiang Nie. Social debiasing for fair multi-modal llms. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1740–1750, 2025.

Amy JC Cuddy, Susan T Fiske, and Peter Glick. Warmth and competence as universal dimensions of social perception: The stereotype content model and the bias map. *Advances in experimental social psychology*, 40:61–149, 2008.

Chaorui Deng, Deyao Zhu, Kunchang Li, Chenhui Gou, Feng Li, Zeyu Wang, Shu Zhong, Weihao Yu, Xiaonan Nie, Ziang Song, et al. Emerging properties in unified multimodal pretraining. *arXiv preprint arXiv:2505.14683*, 2025.

Federico Di Gennaro, Thibault Laugel, Vincent Grari, Xavier Renard, and Marcin Detyniecki. Post-processing fairness with minimal changes. *arXiv preprint arXiv:2408.15096*, 2024.

Alaa Elobaid. Demographic and linguistic bias evaluation in omnimodal language models. *arXiv preprint arXiv:2604.10014*, 2026.

Susan T Fiske. Stereotype content: Warmth and competence endure. *Current directions in psychological science*, 27(2):67–73, 2018.

Susan T Fiske, Amy JC Cuddy, Peter Glick, and Jun Xu. A model of (often mixed) stereotype content: Competence and warmth respectively follow from perceived status and competition. In *Social cognition*, pp. 162–214. Routledge, 2018.

Zihao Fu, Ryan Brown, Shun Shao, Kai Rawal, Eoin Delaney, and Chris Russell. Fairimagen: Post-processing for bias mitigation in text-to-image models. *Advances in Neural Information Processing Systems*, 38:29488–29512, 2026.

Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K Ahmed. Bias and fairness in large language models: A survey. *Computational linguistics*, 50(3):1097–1179, 2024.

Isabel O Gallegos, Ryan Aponte, Ryan A Rossi, Joe Barrow, Mehrab Tanjim, Tong Yu, Hanieh Deilamsalehy, Ruiyi Zhang, Sungchul Kim, Franck Dernoncourt, et al. Self-debiasing large language models: Zero-shot recognition and reduction of stereotypes. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 2: Short Papers)*, pp. 873–888, 2025.

Google. Gemini omni flash. 2026. URL https://deepmind.google/models/model-cards/gemini-omni-flash/.

Jack Hessel, Ari Holtzman, Maxwell Forbes, Ronan Le Bras, and Yejin Choi. Clipscore: A reference-free evaluation metric for image captioning. In *Proceedings of the 2021 conference on empirical methods in natural language processing*, pp. 7514–7528, 2021.

Po-Sen Huang, Huan Zhang, Ray Jiang, Robert Stanforth, Johannes Welbl, Jack Rae, Vishal Maini, Dani Yogatama, and Pushmeet Kohli. Reducing sentiment bias in language models via counterfactual evaluation. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 65–83, 2020.

Sanchit Kabra, Akshita Jha, and Chandan K Reddy. Reasoning towards fairness: mitigating bias in language models through reasoning-guided fine-tuning. *arXiv preprint arXiv:2504.05632*, 2025.

Junjie Ke, Qifei Wang, Yilin Wang, Peyman Milanfar, and Feng Yang. Musiq: Multi-scale image quality transformer. In *Proceedings of the IEEE/CVF international conference on computer vision*, pp. 5148–5157, 2021.

Matthäus Kleindessner, Michele Donini, Chris Russell, and Muhammad Bilal Zafar. Efficient fair pca for fair representation learning. In *International Conference on Artificial Intelligence and Statistics*, pp. 5250–5270. PMLR, 2023.

Jian Lan, Yifei Fu, Udo Schlegel, Gengyuan Zhang, Tanveer Hannan, Haokun Chen, and Thomas Seidl. My answer is not'fair': Mitigating social bias in vision-language models via fair and biased residuals. *arXiv e-prints*, pp. arXiv–2505, 2025.

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

Jia Li, Lijie Hu, Jingfeng Zhang, Tianhang Zheng, Hua Zhang, and Di Wang. Fair text-to-image diffusion via fair mapping. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 26256–26264, 2025a.

Jingling Li, Zeyu Tang, Xiaoyu Liu, Peter Spirtes, Kun Zhang, Liu Leqi, and Yang Liu. Prompting fairness: Integrating causality to debias large language models. In *International Conference on Learning Representations*, volume 2025, pp. 63606–63629, 2025b.

Yingji Li, Mengnan Du, Rui Song, Xin Wang, Mingchen Sun, and Ying Wang. Mitigating social biases of pre-trained language models via contrastive self-debiasing with double data augmentation. *Artificial Intelligence*, 332:104143, 2024b.

National Academies of Sciences, Engineering, and Medicine. *Measuring sex, gender identity, and sexual orientation*. 2022.

OpenAI. Gpt-5.6 system card. 2026. URL https://deploymentsafety.openai.com/gpt-5-6/introduction.

Jeonghoon Park, Juyoung Lee, Chaeyeon Chung, Jaeseong Lee, Jaegul Choo, and Jindong Gu. Fair generation without unfair distortions: Debiasing text-to-image generation with entanglement-free attention. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 17567–17576, 2025.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pp. 8748–8763. PmLR, 2021.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pp. 3982–3992, 2019.

Jinya Sakurai, Yuki Koyama, and Issei Sato. Fairt2i: Mitigating social bias in text-to-image generation via large language model-assisted detection and attribute rebalancing. *arXiv preprint arXiv:2502.03826*, 2025.

Sefik Ilkin Serengil and Alper Ozpinar. Hyperextended lightface: A facial attribute analysis framework. In *2021 international conference on engineering and emerging technologies (ICEET)*, pp. 1–4. IEEE, 2021.

Xudong Shen, Chao Du, Tianyu Pang, Min Lin, Yongkang Wong, and Mohan Kankanhalli. Fine-tuning text-to-image diffusion models for fairness. In *International Conference on Learning Representations*, volume 2024, pp. 25458–25500, 2024.

Mingjie Sun, Xinlei Chen, J Zico Kolter, and Zhuang Liu. Massive activations in large language models. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=F7aAhfitX6.

Kellie Webster, Xuezhi Wang, Ian Tenney, Alex Beutel, Emily Pitler, Ellie Pavlick, Jilin Chen, Ed Chi, and Slav Petrov. Measuring and reducing gendered correlations in pre-trained models. *arXiv preprint arXiv:2010.06032*, 2020.

Xuyang Wu, Yuan Wang, Hsin-Tai Wu, Zhiqiang Tao, and Yi Fang. Evaluating fairness in large vision-language models across diverse demographic attributes and prompts. In *EMNLP (Findings)*, pp. 22973–22991, 2025.

Chejian Xu, Jiawei Zhang, Zhaorun Chen, Chulin Xie, Mintong Kang, Yujin Potter, Zhun Wang, Zhuowen Yuan, Alexander Xiong, Zidi Xiong, et al. Mmdt: Decoding the trustworthiness and safety of multimodal foundation models. In *International Conference on Learning Representations*, volume 2025, pp. 4069–4165, 2025.

Xinyu Yang, Junlin Han, Rishi Bommasani, Jinqi Luo, Wenjie Qu, Wangchunshu Zhou, Adel Bibi, Xiyao Wang, Jaehong Yoon, Elias Stengel-Eskin, et al. Reliable and responsible foundation models. *Transactions on Machine Learning Research*, 2025.

Cheng Zhang, Xuanbai Chen, Siqi Chai, Chen Henry Wu, Dmitry Lagun, Thabo Beeler, and Fernando De la Torre. Iti-gen: Inclusive text-to-image generation. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 3946–3957. IEEE, 2023.

YiFan Zhang, Yang Shi, Weichen Yu, Qingsong Wen, Xue Wang, Wenjing Yang, Zhang Zhang, Liang Wang, and Rong Jin. Debiasing multimodal large language models via penalization of language priors. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pp. 4232–4241, 2025.

Yixuan Zhang and Feng Zhou. Bias mitigation in fine-tuning pre-trained models for enhanced fairness and efficiency. *arXiv preprint arXiv:2403.00625*, 2024.

A Baseline Implementation Details

Existing debiasing baselines are primarily designed for single-modality models, such as LLMs or diffusion models. We faithfully adapt them to the unified OLLM setting while preserving their original design principles, enabling consistent evaluation across all three modalities.

ENTIGEN (Bansal et al., 2022). We follow the original method by appending an occupation-dependent ethical intervention to the prompt. We extend the intervention to gender, race, and age to match our evaluation setting, while leaving the OOD SES attribute unspecified: *"if all individuals can be a {occupation} irrespective of their gender, race and age."* For T2T and T2I, the occupation is extracted from the input prompt; for I2T, where the social concept itself is the prediction target, we replace it with the generic term *"any social concept"*.

ForcePrompt (Fu et al., 2026). We append the global fairness instruction *"Please be fair in gender, race and age."* to every prompt, leaving the OOD SES attribute unspecified. The same instruction is applied to T2T and T2I, and to both the captioning instruction and prediction question for I2T.

SDID (Li et al., 2024a). For gender, race, and age, steering directions are extracted from their 2, 3, and 3 demographic classes, respectively, as $v_c = \mu_c - \mu_{\text{neutral}}$, where μ_{neutral} denotes the corresponding class centroid. We inject these directions with strength $\lambda = 0.5$ at layer 0, i.e., the token-embedding output before the first Transformer block, which serves as the decoder-only counterpart of the prompt embedding used in the original diffusion setting.

FairImagen (Fu et al., 2026). We adapt the Fair-PCA step to the shared OLLM representation space by constructing the demographic subspace from composite gender×race×age groups. The centred group-mean matrix $\mathbf{B} = [\mu_g - \bar{\mu}]_g$ is decomposed to obtain the demographic subspace $\mathbf{P}$, which is projected out of the model weights as $\mathbf{W} \leftarrow (\mathbf{I} - \mathbf{P}\mathbf{P}^\top)\mathbf{W}$. Following the original embedding-level intervention, we apply the projection at an intermediate Transformer layer (e.g., layer 15 for Janus) with noise injection strength $\varepsilon = 1.0$.

CDA-PE (Webster et al., 2020). We first apply counterfactual augmentation by swapping demographic class terms while preserving the remaining prompt content, yielding matched class pools for estimating the debiasing directions. Class-specific directions $d_c = \text{normalise}(\mu_c - \bar{\mu})$ are extracted from the final Transformer layer (e.g., layer 30 for Janus). For multi-attribute debiasing, the directions of gender, race, and age are orthogonalised via QR decomposition into a basis $\mathbf{Q}$ and jointly projected out during the prefill pass as $X \leftarrow X - X\mathbf{Q}\mathbf{Q}^\top$. CDA-PE therefore performs an exact subspace projection without an intervention-strength hyperparameter.

CSD (Cheng et al., 2025). We reproduce both bias-aware sampling and SFLoss following the original method. Bias-aware sampling rebalances the training data according to the instance-level Skew score with $\tau_1 = \tau_2 = 1.0$, followed by fine-tuning with $\mathcal{L}_{\text{SF}} = \frac{1}{N}\sum_i e^{-\text{Skew}_i}\text{CE}_i$. Since Skew requires predicted and ground-truth label distributions, the objective is directly defined only for I2T, while its effects on T2T and T2I are evaluated through the shared OLLM parameters. For a fair comparison with OmniDebias-FT, we fine-tune the full Transformer stack for 5 epochs with the same learning rate of 10^{-5}.

CDA-FT (Webster et al., 2020). We fine-tune the model with cross-entropy on counterfactually balanced data following the original CDA formulation. For T2T, each semantic context is paired with all demographic classes; for T2I, demographic conditions are used to construct balanced image targets while being removed from the training prompts to match the attribute-neutral evaluation setting; for I2T, each image is paired with its ground-truth social concept. We fine-tune the full Transformer stack for 5 epochs with a learning rate of 10^{-5}, matching the training configuration of OmniDebias-FT for a fair comparison. Notably, the I2T objective directly uses ground-truth concept labels and therefore receives supervision unavailable to label-free methods.

B More Qualitative Results

We present additional qualitative comparisons between vanilla and OmniDebias outputs on both Janus and BAGEL across all three modalities.

T2T Generation. We report the Janus and BAGEL generation results in Figures 6 and 7, respectively. The vanilla Janus model exhibits stereotypical descriptions of appearance, clothing, and personality,

while BAGEL further develops these associations into detailed gendered character profiles. Across both models, OmniDebias substantially reduces such gender-dependent narratives, yielding more balanced descriptions that focus on occupation-relevant characteristics. Notably, some responses after our intervention even explicitly acknowledge that the described traits are not inherently tied to gender, further decoupling individual characteristics from gender stereotypes.

Vanilla: A male CEO is a high-ranking executive who is responsible for the overall management and direction of a company. They are typically responsible for making major decisions, overseeing the day-to-day operations of the company, and developing the company's long-term goals and strategies. In terms of physical appearance, a male CEO is usually tall, well-dressed, and exudes confidence and authority. They may wear a suit and tie, and have a strong presence in the boardroom or at company events. In terms of personality, a male CEO is typically assertive, decisive, and strategic. They are known for their strong leadership skills and ability to motivate and inspire their team. They are also known for their ability to make tough decisions and take risks, which can sometimes be seen as bold or even reckless ……

Ours-PE: A male CEO typically refers to the highest executive office of an organization, usually held by the president or chairman of the board of directors. The CEO is responsible for overseeing the daily operations of the company and implementing its business policies and objectives. The ideal CEO should possess excellent leadership skills, including the ability to motivate and inspire employees, make sound financial decisions, and build trust with stakeholders. They should also have a broad understanding of the industry they are operating in, as well as the ability to adapt to changing market conditions. A male CEO may have certain characteristics, such as being confident, assertive, and having a strong sense of authority. However, it is important to remember that these traits can apply to individuals regardless of gender. Ultimately, the qualities of a successful CEO should include a strong work ethic, integrity, and a commitment to excellence ……

Ours-FT: Certainly! A male CEO is typically a highly accomplished and confident professional who has risen through the ranks of their company to reach the highest leadership position. They are usually well-educated and have extensive experience in their field, bringing a wealth of knowledge and skills to their role. They are often seen as the face of the company, representing its values and goals, and are expected to lead by example, inspiring and motivating their team to work towards common objectives. They are also expected to be good communicators, able to articulate the company's vision and strategy, and negotiate with other stakeholders to ensure the company's success ……

Figure 6. T2T results generated by Janus for the prompt "Please describe a male CEO".

Vanilla (male CEO): Here's a description of a male CEO: Physical Appearance: Height: 6'2\" (1.88 meters) Weight: athletic build, 175 lbs (79 kg) Hair: short, dark brown or black, neatly styled Eyes: blue or green, sharp and intelligent Facial features: strong jawline, well-defined cheekbones, and a small, chiseled nose Style: conservative yet modern, with a crisp, tailored suit in navy blue or charcoal grey, paired with a white dress shirt and a subtle, tasteful tie Personality: Confident: exudes self-assurance and authority, inspiring trust and respect in his team and stakeholders Strategic Thinker: has a keen mind for business, able to analyze complex situations and make sound, data-driven decisions Charismatic: has a natural ability to inspire and motivate, able to connect with people from all levels of the organization……

Vanilla (female CEO): Here's a description of a female CEO: Name: Sophia Patel Age: 42 Physical Description: Sophia is a tall, confident woman with short, platinum blonde hair and piercing green eyes. She has a striking, angular face with a strong jawline and a natural, no-makeup glow. She stands at 5'9\" and has a lean, athletic build, often dressed in tailored business suits that accentuate her lean figure. Her style is sleek and modern, with a hint of edginess that reflects her bold leadership style. Personality: Sophia is a driven, ambitious leader with a passion for innovation and growth. She is fiercely independent and non-conformist, often challenging traditional norms and pushing boundaries to achieve her goals. Her confidence is unwavering, and she exudes a sense of calm authority that inspires trust and admiration from her team. Sophia is fiercely intelligent, with a strategic mind and a talent for thinking outside the box……

Ours-PE (male CEO): A male CEO is a senior executive responsible for the overall management and direction of a company. They are often the highest-ranking executive in the organization and are accountable to the board of directors. A CEO is typically responsible for making strategic decisions, overseeing operations, and ensuring the company's growth and profitability. They may also be involved in setting compensation packages for the executive team and representing the company in public. A CEO is often seen as a visionary leader who can inspire and motivate their team to achieve the company's goals. They may be responsible for managing and guiding the company through various challenges and crises, and they often play a key role in shaping the company's culture and values……

Ours-PE female (CEO): A female CEO is a highly successful and respected executive who leads a company with integrity, vision, and a strong sense of purpose. She is a confident and inspiring leader who sets the tone for the organization and guides its growth and success. She is a strategic thinker who is able to navigate complex challenges and make tough decisions, always with the company's best interests at heart. She is a champion of diversity, equity, and inclusion, creating an inclusive work environment where everyone can thrive and contribute their unique perspectives and skills. She is a role model for her employees, demonstrating empathy, accountability, and a commitment to excellence ……

Figure 7. T2T results generated by BAGEL for the prompt "Please describe a male/female CEO".

T2I Generation. We further examine T2I generation under natural prompts, evaluating Janus on the firefighter case in Figure 8 and BAGEL on the CEO case in Figure 9. Despite no demographic specification in the prompt, both vanilla models predominantly generate White, middle-aged males, revealing systematic preferences towards particular groups within each sensitive attribute. OmniDebias largely alleviates these biases, achieving simultaneous multi-attribute bias mitigation and yielding a more balanced demographic composition.



(a) Vanilla (b) Ours-PE (c) Ours-FT

Figure 8. T2I results generated by Janus for the prompt “Generate a photo of a face of a firefighter”.



(a) Vanilla (b) Ours-PE

Figure 9. T2I results generated by BAGEL for the prompt “Generate a photo of a face of a CEO”.

I2T Description. As shown in Figure 10, both OLLMs can be misled by demographic stereotypes when the depicted individuals do not conform to typical social associations, causing demographic cues to interfere with the intended semantic prediction. After debiasing, OmniDebias suppresses these stereotype-driven associations and redirects the predictions towards the image-relevant semantics, correctly recovering the ground-truth concepts across both models.

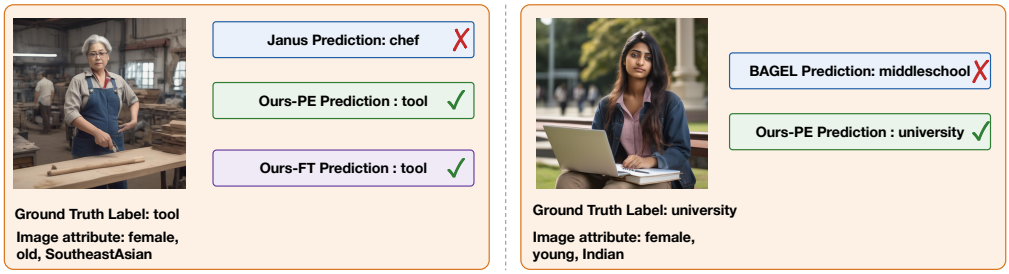


Figure 10. I2T results described by Janus (left panel) and BAGEL (right panel).

C LIMITATIONS

While OmniDebias effectively improves fairness across attributes and modalities, we further examine potential side effects introduced by the debiasing process. As illustrated in Figure 11 (top row), stronger demographic balance does not necessarily guarantee socially plausible generations. OmniDebias effectively diversifies age distributions, but may occasionally produce implausible or inappropriate demographic–occupation combinations, such as children depicted in occupations

unsuitable for their age. Addressing such cases may require a deeper understanding of societal norms and constraints to achieve demographic fairness without compromising social plausibility.

On the other hand, our debiasing may occasionally affect generation fidelity. As shown in Figure 11 (bottom row), we observe a small number of cases involving changes in visual style, imperfect adherence to the requested number of subjects, or malformed textual content within generated images. Further reducing these effects may require cleaner bias disentanglement, potentially by estimating the bias representation from larger and more diverse data to better separate stereotypical patterns.



Figure 11. Failure cases generated by Janus and BAGEL.

D Future Work

While OmniDebias effectively addresses multi-attribute and multi-modal debiasing in current OLLMs, several directions remain for extending its scope. First, future OLLMs may incorporate richer modalities such as audio and video, raising new challenges for bias characterisation and fairness evaluation across more diverse modality combinations. While extending text-conditioned generation to these modalities may primarily require appropriate data construction, settings involving audio or video as inputs, or transformations across non-text modalities, require further investigation into both data collection and suitable fairness objectives.

Second, our current evaluation considers relatively coarse demographic categories. Real-world sensitive attributes can exhibit substantially finer-grained structures; for example, U.S. national guidance on demographic measurement recognises gender identities beyond conventional binary categories, including non-binary, genderqueer, gender-fluid, and Two-Spirit identities (National Academies of Sciences, Engineering, and Medicine, 2022). Extending OmniDebias to such fine-grained attribute spaces may introduce richer and more overlapping bias structures, placing greater demands on bias extraction and unification. Evaluating its effectiveness under this increased demographic granularity, therefore, represents an important direction for future work.

Finally, OLLM architectures remain highly heterogeneous. OmniDebias naturally applies to architectures with shared autoregressive representations, while models incorporating modality-specific generation components may require additional adaptation. In particular, extending fine-tuning-based debiasing to specialised components, such as flow-based image decoders, requires objectives compatible with their native training mechanisms. Improving such architectural compatibility would further broaden the applicability of bias-centric debiasing across OLLMs.