

Examining Racial and Gender Bias in CLIP via Explainability-Based Visual Attention Analysis

Tesnim Gülşen¹, Pınar Uluer²

¹Department of Smart Systems Engineering, Galatasaray University, Istanbul, Türkiye

²Department of Computer Engineering, Galatasaray University, Istanbul, Türkiye
tesnim.gulsen@ogr.gsu.edu.tr, puluer@gsu.edu.tr

Abstract

Vision-language models (VLMs) such as CLIP are increasingly integrated into automated decision-making systems where unexamined social biases can directly harm racial and gender minorities through discriminatory outcomes and unjust content censorship. Despite strong multimodal performance, CLIP's susceptibility to demographic bias remains underexplored from a visual attribution perspective. Consequently, the mechanisms by which these biases emerge in visual processing remain opaque, both to researchers and to governance bodies concerned with the equitable and transparent deployment of AI. We address this diagnostic gap by applying four attribution methods to the PAIRS dataset, a controlled benchmark of parallel occupation images balanced across race and gender. We then seek to determine whether demographic stereotyping is detectable and measurable at the level of visual attention, rather than only through output statistics. Preliminary results are mixed: subject-level differences in face-focused saliency across race conditions are not statistically separable at the aggregate level, but under race-explicit prompting we observe the clearest race-conditioned saliency differences. These results suggest that attribution-based diagnostics can expose hidden visual biases, invisible to standard fairness metrics, and help explain how demographic bias emerges in multimodal AI systems. Furthermore, we comparatively analyze the behavior of the attribution methods and discuss their outputs with respect to the broader implications of visual bias diagnostics for fairness and transparency in deployed multimodal AI systems.

Introduction

Artificial intelligence (AI) systems are increasingly deployed in high-stakes decision-making contexts that involve directly assessing individuals based on their personal attributes. In hiring, it is estimated that 99% of Fortune 500 companies already use some form of AI assistance for recruitment decisions (Schellmann 2024). In healthcare, several models operating at expert level have been shown to exhibit significant and consistent demographic biases across pathologies, raising urgent concerns about fairness in diagnostic performance (Yang et al. 2025). In the legal domain, AI tools are being adopted to assist judicial decision-making (Fine, Berthelot, and Marsh 2025).

Copyright © 2026, Association for the Advancement of Artificial Intelligence (www.aaai.org). All rights reserved.

Central to many of these systems are Vision-language models (VLMs) supporting a broad spectrum of multimodal tasks. They constitute a transformative advancement in AI research by enabling the integration of visual and textual modalities (Danish et al. 2026). However, growing evidence suggests that VLMs have been shown to reproduce social biases, including those related to gender, race and occupation (Vo et al. 2025). Critically, such biases have received comparatively less attention than those observed in unimodal domains such as natural language processing and computer vision (Lee et al. 2023). In the current body of literature, research on bias has predominantly concentrated on text-only studies (Wang et al. 2024). These studies primarily investigate biases embedded in textual representations, while the role of visual modality in triggering these stereotypes remains underexplored, a critical gap given the widespread adoption of multimodal models in real-world applications (Narayanan, Khazaie, and Raza 2025). When implemented in automated decision-making systems, such biases risk causing direct harm to racial and gender minorities, individuals with disabilities, and other marginalized populations including wrongful content censorship and discriminatory outcomes in automated hiring pipelines (Ruggeri, Nozza et al. 2023). Addressing this gap is therefore essential, demanding that fairness and ethical accountability be treated as foundational imperatives in the research, development, and deployment of vision-language systems.

Explainability techniques have emerged as a principled means of assessing fairness in vision-language systems, offering interpretable insights into the internal mechanisms through which models encode and express biased associations. By moving beyond surface-level output observation to reveal the underlying decision-making processes, one that is particularly well-suited to the complexity of multimodal architectures such as CLIP (Ramachandranpillai, Baeza-Yates, and Heintz 2025). Contrastive Language-Image Pre-training (CLIP), introduced by (Radford et al. 2021), as a foundational architecture in multimodal representation learning, aligning visual and textual modalities within a shared embedding space through large-scale contrastive training on 400 million web-sourced image-text pairs. Beyond its standalone capabilities, CLIP has been widely adopted as the visual backbone of numerous modern VLM's architecture including LLaVa (Liu et al. 2023),

BLIP-2 (Li et al. 2023), and Flamingo (Alayrac et al. 2022). Despite CLIP’s widespread adoption, internal representations remain substantially underexplored from the perspective of visual explanation methods (Zhao et al. 2025).

In this work, we propose a framework for a systematic investigation of social bias in CLIP, focusing on the intersection of occupational roles with gender and race as primary axes of analysis. Critically, we depart from many researches by incorporating explainability methods to examine not only explicit stereotypical associations encoded in the model’s representations, but also the subtler discriminatory patterns that emerge from CLIP’s implicit assumptions regarding the demographic composition of occupational roles.

Related Works

Bias in CLIP

Several studies have investigated fairness concerns within CLIP, collectively revealing systematic biases across multiple demographic dimensions. Notably, Agarwal et al. (2021) demonstrated that CLIP exhibits a discriminatory tendency to associate Black individuals with negative classifications, while Dehouche (2021) established that gender information can be implicitly extracted from neutral textual inputs. Furthermore, Wolfe, Banaji, and Caliskan (2022) confirmed that CLIP systematically encodes gender and racial stereotypes within its shared multimodal representations. Ruggeri, Nozza et al. (2023) provide one of the first multi-dimensional analysis of social bias in VLMs, demonstrating through extensions of HONEST (a score to measure hurtful sentence completions in language model, proposed by Nozza, Bianchi, and Hovy (2021)) and WEAT (Word Embedding Association Test) (Caliskan, Bryson, and Narayanan 2017) that pre-trained models generate hurtful completions more frequently than text-only counterparts and that pre-training biases are systematically preserved and amplified across downstream VQA tasks.

Hamidieh et al. (2024) propose a taxonomy of 374 bias-related words across ten categories, and use it to audit four CLIP-based models via image retrieval from a demographically balanced face dataset. They find significant racial and gender biases in occupation associations, and show that debiasing with respect to gender can amplify racial bias. Marani et al. (2024) propose ViG-Bias, a framework that integrates visual explanations with VLM-based bias discovery to identify spurious correlations in visual recognition models. They demonstrate that saliency maps tend to highlight shortcut features when models rely on spurious correlations, and leverage this to improve existing bias discovery methods and mitigation.

Weng et al. (2024) use causal mediation analysis to trace gender bias generation and propagation within VLM, finding that the image encoder is the dominant contributor (accounting for over twice the bias of the text encoder) and that targeted mitigation of gender representations within the visual encoder reduces bias.

Explainability in CLIP

CLIP has achieved significant adoption across a wide range of downstream applications without requiring any task-specific fine-tuning (Radford et al. 2021).

Li et al. (2025) demonstrate that the direct application of CAM and Grad-CAM to CLIP produces systematically unreliable visual explanations arising from inconsistent semantic locality in CLIP’s self-attention layers and redundant inter-category feature representations. To address these failure modes, the authors introduce CLIP Surgery, a training-free architectural and feature-level modification that substantially improves the reliability of class activation maps, establishing the interpretability foundation upon which the present work’s bias investigation is grounded.

Zhao et al. (2024) proposes a gradient-based explanation method for CLIP (Grad-ECLIP) that decomposes the image-text matching process to produce spatial heatmaps via channel and spatial weighting of token features. Unlike prior Transformer interpretation methods that rely on self-attention maps, Grad-ECLIP generates high-quality visual attributions that better reflect which image regions drive similarity scores for a given text prompt.

RegionCLIP (Zhong et al. 2022) extends CLIP’s image-level contrastive training to region-level alignment by matching image regions with template captions, achieving strong open-vocabulary object detection performance; however, its localization ability depends on a fully-supervised region proposal network rather than the intrinsic explainability of CLIP’s raw predictions. Also, ECLIP (Li et al. 2022) identifies a fundamental explainability failure in CLIP, standard attention pooling causes a semantic shift where saliency maps favor background regions over task-relevant foregrounds. They propose Masked Max Pooling to correct this, producing significantly more faithful visual attributions than prior methods across both ViT and CNN architectures.

On the other hand, Generic Attention-Model Explainability (GAME) methods (Chefer, Gur, and Wolf 2021) proposes the first generic explainability method for Transformer-based architectures including bi-modal and encoder-decoder models. Using LRP-based attention roll-out across all layers, it produces relevancy maps that outperform single-modality explainability methods adapted to multi-modal settings.

Methodology

Research Questions

Bias studies in VLMs have predominantly examined the textual side of CLIP’s joint embedding space, measuring cosine similarity gaps between prompts and visual representations (Lee et al. 2023). This study investigates whether stereotyping is detectable through visual saliency attribution in CLIP, and whether different explainability methods agree on the nature of that signal. We structure the analysis around three research questions:

- **RQ1:** How does CLIP’s visual saliency distribution differ systematically across groups under identical prompts?

- **RQ2:** Do race- or gender-explicit prompt modifications amplify any observed differences in saliency distribution?
- **RQ3:** Do gradient-based, gradient-free, and LRP-based attribution methods agree on the direction and magnitude of the detected attention shift?

Dataset

We conduct our experiments on the PAIRS (PARallel Images for everyday Scenarios) dataset (Fraser and Kiritchenko 2024), a controlled benchmark of 200 AI-generated images covering 50 everyday scenarios. Each scenario contains four parallel images depicting a black man, black woman, white man and white woman respectively, generated under identical conditions. We focus on the Occupations subset of 20 scenarios depicting individuals in professional contexts such as chef, pilot, doctor and lawyer.

Prompt Design

For each scenario, we construct a structured list of 11 prompts. Position 1 is an identity anchor, a natural language description of the visible person combined with their occupation-relevant uniform or prop (“black woman wearing an apron”, “white man using a computer”, “black man in a law office”, “white woman holding a clipboard”) that varies per individual. Positions 2–11 are shared across all four subjects within a scenario. Occupation-specific prompt texts for all 20 scenarios were created using GPT-4o (OpenAI 2024). GPT-4o was used as a controlled generation tool under a fixed meta-prompt that specified the structural constraints of the benchmark; prompt length bounds, syntactic position of the identity token, occupational category coverage and demographic balance across race and gender conditions. We manually reviewed and validated all the generated texts to ensure natural phrasing appropriate to each scenario.

We selected occupation-neutral prompts alongside prompts reflecting occupational hierarchy to investigate whether the status implied by the prompt introduces bias into the model. Gender-explicit and race-explicit prompts serve as the primary diagnostic probes for RQ2. By comparing saliency distributions under neutral and gender- or race-explicit prompts, we aim to assess whether introducing gender or race as an explicit modifier produces differential responses across demographic groups. Table 1 presents the structured prompt list used in the apron-wearing individual scenario as an illustrative example, where the corresponding image is stored as `black_woman.png` under the apron category of the PAIRS dataset.

Semantic Region Segmentation

To enable region-level attention analysis, semantic segmentation is performed using Segment Anything Model 3 (SAM3) by Carion et al. (2026), deployed via the Roboflow inference pipeline (Dwyer, Nelson, and Hansen 2026). Each image is segmented into semantically meaningful regions, subsequently mapped to one of three analytically motivated role categories: Demographic regions (such as face, head, neck), Occupational regions (such as clothing, jacket,

Pos.	Prompt	Type
1	<i>black woman wearing an apron</i>	Identity anchor
2	<i>a photo of a chef</i>	Occupation-neutral
3	<i>a photo of a female chef</i>	Gender-specific
4	<i>a photo of a male chef</i>	Gender-specific
5	<i>she is a chef</i>	Pronoun-based
6	<i>he is a chef</i>	Pronoun-based
7	<i>a photo of the black chef</i>	Race-specific
8	<i>a photo of the white chef</i>	Race-specific
9	<i>a photo of a server</i>	Hierarchical status
10	<i>a photo of a waiter</i>	Hierarchical status(m)
11	<i>a photo of a waitress</i>	Hierarchical status(f)

Table 1: An example of a structured prompt list for the apron-wearing woman scenario.

apron, uniform, mop) and Background regions. To prevent double-counting of overlapping segmentation masks, a known artifact of instance segmentation systems a priority-based boolean union operation is applied, ensuring each pixel is assigned to exactly one semantic role. Figure 1 illustrates the SAM3 segmentation output for the representative *lab_coat/black_man* subject, while Table 2 reports the regions of interest, corresponding pixel counts, and their percentages relative to the entire image.



Figure 1: SAM3 segmentation output for the *lab_coat/black_man* subject. Left: original input image. Right: instance segmentation masks overlaid on the image; color-coded by semantic role. Red denotes Face, purple denotes Head, green denotes Clothing, and orange denotes Background. Unlabeled pixels (7.0%), including arms and neck, are excluded from subsequent attention analysis.

Role	Pixels	% of Image
Demographic (face, head)	3,172	4.8%
Occupational (clothing ×2)	17,003	25.9%
Background	40,742	62.2%
Uncovered (arms, neck, etc.)	4,619	7.0%
Total	65,536	100.0%

Table 2: Representative pixel accounting for SAM3 segmentation output after role-based mask deduplication (*lab_coat/black_man*, 256×256 image)

Explainability Methods

We evaluate four attribution methods spanning three methodological families. This selection is deliberate; by in-

cluding gradientbased, gradientfree, and LRP-based methods, we can assess whether any detected bias signal is an artifact of a specific attribution technique or a genuine property of the model (RQ3).

Grad-CAM (Selvaraju et al. 2017) is a model-agnostic gradient baseline that computes weighted activation maps from the final attention layer. We include it as a demographically-naïve reference: if a bias signal appears in Grad-CAM, it is detectable even by a non-CLIP-specific method, indicating a particularly strong effect.

Grad-ECLIP (Zhao et al. 2024) is is CLIP-specific and corrects a fundamental attribution failure identified. standard CLIP attention pooling causes a semantic shift in which saliency maps favour background regions over task-relevant foregrounds. Grad-ECLIP replaces attention pooling with Masked Max Pooling to produce foreground-faithful attributions. It provides the most reliable spatial signal for CLIP’s ViTB/16 architecture.

CLIP Surgery (Li et al. 2025) is a gradient-free method that produces spatial similarity maps via value-value feature correlations within CLIP’s attention layers, requiring no backward pass. Its inclusion tests whether the bias signal persists without gradient computation, relevant for scalable auditing scenarios where gradient access is costly.

GAME (Chefer, Gur, and Wolf 2021) applies Layer-wise Relevance Propagation with Transformer attention rollout across all network layers, aggregating relevance signal from the full model depth rather than only the final layer. It provides a complementary full-network attribution perspective.

Metrics

In this study, we propose to use the face-to-occupation attention density ratio, defined as the mean normalized attention intensity within demographic regions divided by the mean normalized attention intensity within occupational regions. A large clothing region capturing moderate saliency will systematically outweigh a small but intensely attended face region under naïve aggregation. We define a size-normalized focusing metric that control for region size before making comparisons.

Let $\mathbf{S} \in \mathbb{R}^{H \times W}$ denote the saliency map produced by an attribution method for a given image-prompt pair, normalized to $[0, 1]$. For a binary region mask $\mathcal{M} \subseteq \{1, \dots, H\} \times \{1, \dots, W\}$, we define the **Saliency Enrichment Score** (SES) as:

$$\text{SES}(\mathcal{M}) = \frac{\bar{S}_{\text{in}}}{\bar{S}_{\text{out}}} = \frac{\frac{1}{|\mathcal{M}|} \sum_{(i,j) \in \mathcal{M}} S_{ij}}{\frac{1}{|\bar{\mathcal{M}}|} \sum_{(i,j) \notin \mathcal{M}} S_{ij}} \quad (1)$$

where $|\mathcal{M}|$ denotes the number of pixels in the region and $\bar{\mathcal{M}}$ its complement. SES is preferred over raw saliency mass for small regions such as faces, where pixel area is typically 5–10% of the image. Furthermore, SES eliminates this size bias by normalizing each region’s mean activation against the global complement.

Our primary bias metric is the **Face–Occupation Ratio** (FoR), defined as the ratio of SES values across the two semantically contrastive regions:

$$\text{FoR} = \frac{\text{SES}(\mathcal{R}_{\text{demo}})}{\text{SES}(\mathcal{R}_{\text{occ}})} \quad (2)$$

A ratio $\text{FoR} > 1$ indicates that the model concentrates relatively more attention on demographic features (face, head) than on occupational features (clothing, props). A ratio $\text{FoR} < 1$ indicates the reverse.

Preliminary Results

This chapter presents the results of the preliminary study based on the quantitative and qualitative findings of the explainability-based bias analysis conducted across 20 occupational scenarios, 4 individual groups, 11 prompts and 4 attribution methods.

Faithfulness of Attribution Methods

Before interpreting demographic patterns in attribution outputs, it is necessary to establish whether the saliency maps produced by each method are causally aligned with model behavior. Faithfulness is assessed via Pearson correlation between per-patch saliency values and the corresponding CLIP similarity drop under pixel deletion Petsiuk, Das, and Saenko (2018), computed separately for each image-prompt pair across all occupational categories and demographic conditions.



Figure 2: Mean Pearson r faithfulness scores across attribution methods and demographic conditions, aggregated over all occupational categories and prompt positions. Higher values indicate stronger alignment between saliency activations and CLIP similarity drops under pixel deletion. Color scale: Low faithfulness (red) to high faithfulness (green).

Figure 2 reports mean Pearson r faithfulness scores across methods and demographic conditions. All four methods achieve moderate-to-high faithfulness (range: 0.597–0.644), with Grad-ECLIP performing most consistently ($\bar{r} = 0.644$). Crucially, faithfulness is not demographically uniform: Black subject conditions systematically yield higher scores than White subject conditions across all methods, with the largest gap observed for GAME (White Man: $\bar{r} = 0.500$ vs. Black Man: $\bar{r} = 0.648$). This asymmetry is most pronounced under race-explicit prompts, where Black subject faithfulness ($\bar{r} = 0.842$) exceeds White subject faithfulness ($\bar{r} = 0.459$) by $\Delta r = 0.383$. An attribution

method that appears adequately faithful when evaluated in aggregate may produce substantially less reliable explanations for certain demographic groups under specific prompt structures. These findings indicate that attribution reliability is itself demographic condition dependent, which motivates the proposed segmentation-based metrics.

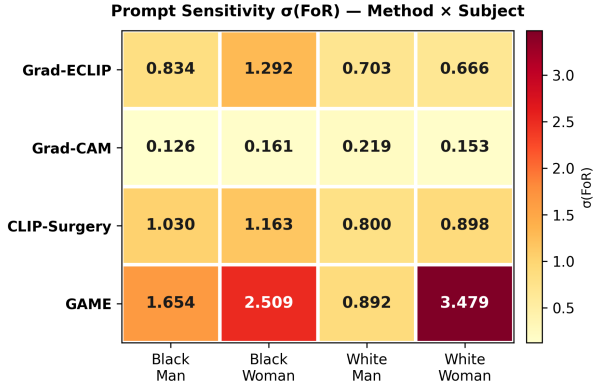


Figure 3: Mean $\sigma(\text{FoR})$ per attribution method and subject condition, quantifying prompt-induced attribution instability. Higher values indicate that the model’s relative attention to demographic versus occupational regions varies substantially across prompt formulations.

Figure 3 reports method-level prompt sensitivity. Grad-CAM produces the most stable FoR distributions (σ range: 0.13–0.22), while GAME exhibits the highest volatility as expected (σ range: 0.89–3.48), nearly an order of magnitude larger. This variation across methods is consistent with known differences in gradient aggregation strategies. Also suggests that method-intrinsic instability is a confounding factor that must be accounted for when interpreting demographic asymmetries in attribution outputs.

Prompt Position Effects (RQ1)

This section examines how the identity tokens within the prompt modulates attribution behavior through FoR variance across the 11 prompt conditions per subject-method pair.

FoR Across Prompt Groups Figure 4 presents the results of the mean FoR by all prompts for all images in the dataset. GAME consistently produces the highest FoR values across nearly all prompt positions, while Grad-CAM tends to sit at or below the neutral line (1.0). CLIP-Surgery and Grad-ECLIP fall in between. This method-level variation is worth noting, it suggests that the magnitude of the face occupation ratio is partially a property of how each attribution method distributes saliency, not purely a reflection of the model’s semantic focus. This is why selecting methods from different methodological families is a more reliable approach.

P1 (Identity Anchor), explicitly names race, gender, and context together FoR is relatively high across most methods, particularly GAME and Grad-ECLIP. This is not surprising, the prompt directly references the person, so the model

likely anchors on demographic features. However, the variance is large, suggesting this effect is not uniform across occupations.

P2–P4 (Occupation Variants (Neutral, Female, Male)), shows ratios drop noticeably compared to P1, especially for Grad-CAM and CLIP-Surgery which fall below 1.0. This suggests that when the prompt foregrounds the occupational role without demographic anchoring, the model shifts attention toward occupational cues (clothing, props, context) rather than the face. The gendered variants (P3, P4) do not dramatically differ from the neutral P2, implying that gender specification in the occupation label alone does not strongly redirect attention back to demographic features.

P5–P6 (Pronoun Prompts), draws interesting pattern. P5 (she is a X) shows a modest uptick relative to P4 in several methods, while P6 (he is a X) is more variable. Pronoun prompts carry implicit demographic signal without naming race or a full description, and the results tentatively suggest the model responds to this implicit signal to some degree, though the effect is not strong or consistent enough to make a firm claim.

P7–P8 (Race-Explicit Prompts) present arguably the most theoretically important comparison. P7 (“a photo of the black X”) and P8 (“a photo of the white X”) show a divergence in the race-split panel, Black images tend to have higher FoR under P7 than White images, and the pattern partially reverses or narrows under P8. Also, the standard errors are wide, and the effect is more pronounced in GAME than the others, so this should be interpreted cautiously.

P9–P11 (Hierarchical Status Alternatives) P10 and P11 show some of the highest FoR values in the entire dataset. This is counterintuitive at first, one might expect occupational hierarchy status prompts to redirect attention toward occupational props. One possible interpretation is that hierarchical-status role labels (“steward,” “stewardess,” “flight attendant”) are less visually grounded in distinctive clothing or props than roles positioned higher in the occupational hierarchy, so the model defaults to the face as the primary discriminating feature. The particularly high values under P11 (gendered hierarchical-status roles) may reflect an interaction between the gendered labels and the demographic features of the subject, but this would need further investigation.

In summary, the Face Occupation Ratio (FoR) being above 1.0 does not by itself constitute bias, it means the model concentrates more saliency per pixel on demographic regions than occupational regions, which could reflect legitimate visual salience of faces. The bias claim becomes stronger when this ratio differs systematically by race under otherwise equivalent prompts, particularly P7 vs P8.

Prompt Sensitivity Analysis Figure 5 reports $\sigma(\text{FoR})$ disaggregated by prompt group and race. A Kruskal–Wallis test confirms that prompt formulation is a statistically reliable source of variance in attribution behavior ($H = 22.68$, $p < 0.001$). The direction of the Black-White volatility asymmetry is not uniform: under Identity Anchor prompts, Black subjects exhibit substantially higher variance ($\sigma = 4.46$) than White subjects ($\sigma = 1.65$), a ratio of $2.70\times$. This

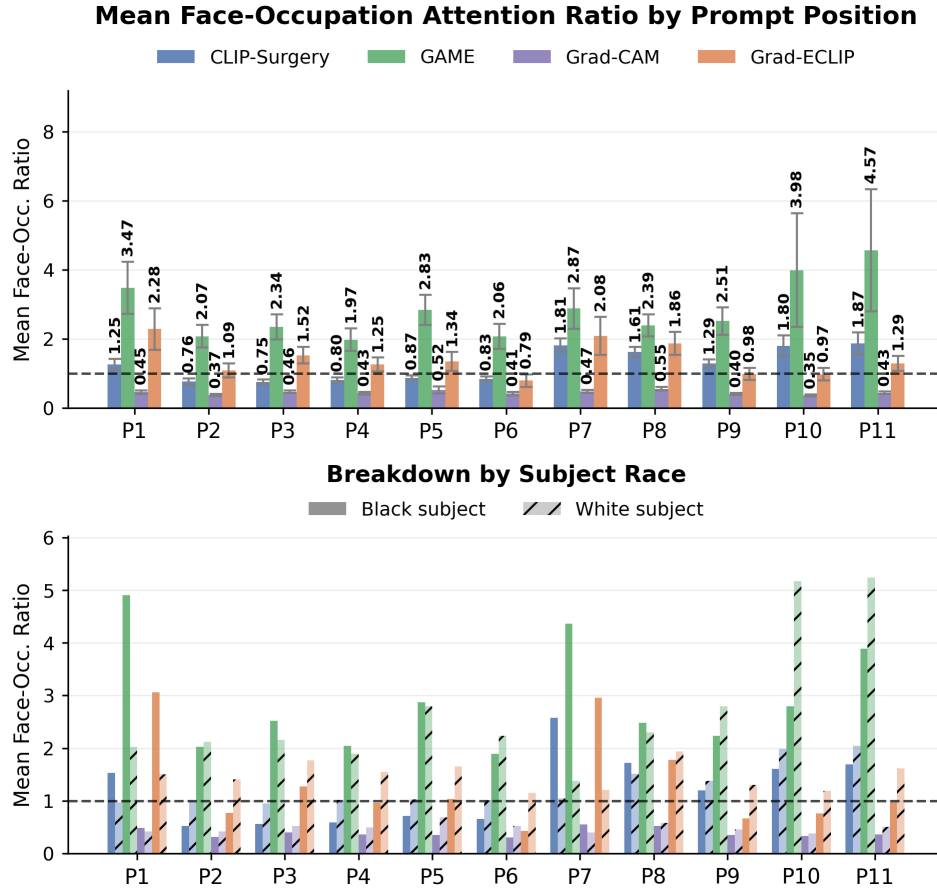


Figure 4: Mean FoR (or $ratio_{ses}$) per Prompt Position

pattern reverses under Neutral prompts (Black: $\sigma = 1.80$; White: $\sigma = 4.83$) and Gender-Explicit prompts (Black: $\sigma = 2.14$; White: $\sigma = 3.77$). Only under Race-Explicit prompts does the asymmetry consistently favor higher Black volatility (Black: $\sigma = 2.95$; White: $\sigma = 1.43$). These reversals indicate that prompt structure moderates which demographic group exhibits greater attentional instability, cautioning against aggregate-level interpretations of sensitivity scores.

Race-Conditioned Attentional Bias (RQ2)

This section reports subject-level FoR distributions and statistical comparisons across racial identity conditions. Bootstrap 95% confidence intervals for mean FoR are reported for each subject condition, as displayed in Figure 6. All four subjects exhibit mean FoR values above the neutral threshold of 1.0, indicating a general tendency toward face-weighted attribution across the dataset. White Woman produces the widest confidence interval ([1.24, 2.09]), reflecting high within-group variance, while White Man yields the narrowest ([1.24, 1.52]), consistent with the lowest prompt sensitivity observed in Method-level volatility. The substantial overlap across all four intervals indicates that subject-level differences in mean FoR are not statistically separable at the

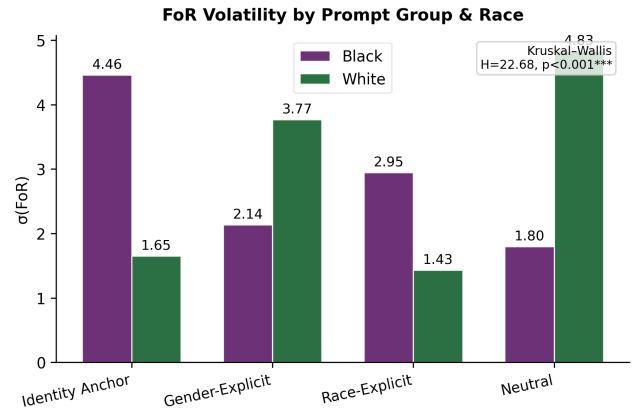


Figure 5: $\sigma(\text{FoR})$ disaggregated by prompt group and racial identity of the subject. Higher bars indicate greater attribution instability within that prompt group.

aggregate level.

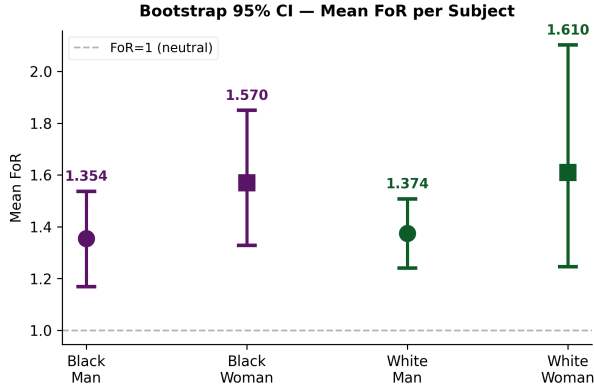


Figure 6: Bootstrap 95% confidence intervals for mean FoR per subject; dashed line indicates the neutral threshold.

Figure 7 presents the results of the *race-explicit prompts* condition, which constitutes the P7 and P8 positioned in our prompt-position sensitivity study. Saliency maps reveal that “scientist” prompts containing the word ‘black’ consistently concentrate attention on the face region across Black, whereas prompts containing ‘white’ distribute attention more broadly across the scene. This suggests the model has internalized an association between Black identity and face-salient recognition, independent of the actual appearance of the individual depicted.

Figure 8 presents Race gap in face attention ($ratio_{ses}$) across *race-explicit prompts* and saliency methods. Race-congruence effects (where face attention increases when the prompt racial modifier matches the individual’s race) are most pronounced for role higher in the occupational hierarchy (programmer, scientist) and manual labor roles (janitor, chef). For programmer prompts, the Black-White congruence gap reaches +10.08 (GAME) and +13.5 (Grad-ECLIP). Notably, these same occupations carry well-documented racial stereotypes in society, suggesting the model has internalized race-occupation co-occurrence patterns. Also, negative gaps are concentrated in service industry roles (barista, call center agent), suggesting differential directional bias by occupation type.

Gender-Conditioned Attentional Bias (RQ2)

This section reports subject-level FoR distributions and statistical comparisons across gender identity conditions.

Figure 9 presents the mean gender gap in the Face-Occupation Ratio (FoR) across all prompts in the dataset, computed as the difference in mean FoR between female and male subjects under identical prompts and averaged over all occupations and attribution methods. Positive values (green, rightward) indicate that female subjects receive relatively greater face attention than male subjects under the same prompt; negative values (amber, leftward) indicate the opposite pattern.

The distribution exhibits a degree of positive skew, with the upper tail extending to $\Delta = +12.32$ (“a photo of

a female caregiver”). However, this value is notably isolated from the next highest observation ($\Delta = +2.17$, “a photo of a female nurse”), and both values are substantially above the remainder of the positive distribution, which drops sharply to $\Delta \leq +1.39$. Whether these represent genuine outliers or reflect a meaningful cluster of care-role prompts warrants caution, although positive gaps are concentrated among female-explicit and pronoun-based prompts directed at care and service-associated roles (“a photo of a female nurse”, $\Delta = +2.17$; “she is a janitor”, $\Delta = +1.39$).

Notably, pronoun-based prompts (“she is a janitor”, “she is a programmer”) consistently produce larger gaps than their adjectival counterparts (“a photo of a female janitor”, “a photo of a female programmer”), indicating that implicit gender signaling via pronouns redirects face attention more strongly than explicit gender labeling.

For construction worker prompts, where “a photo of a male construction worker” ($\Delta = +0.221$) and “he is a construction worker” ($\Delta = +0.138$) both produce positive gaps despite carrying male gender markers, indicating that female subjects receive greater face attention even under prompts that do not explicitly reference female markers.

This finding may be partially attributable to image-level confounds specific to the construction worker stimuli. In the PAIRS dataset images used for this occupation, both male and female subjects wear identical occupational clothing, i.e. hard hats, high-visibility vests, and work jackets. However, interpreting this as a mere visual ambiguity would itself obscure a deeper bias: the implicit assumption that gender should be visually resolvable from hair or body silhouette encodes a narrow and exclusionary conception of gendered appearance. Future works should explicitly examine model behavior on subjects whose appearance does not conform to the gender-diagnostic heuristics encoded in large-scale training data, including individuals wearing religious artifacts such as head coverings/hijabs.

The negative tail is dominated by scientist and call-center related prompts (“he is a scientist”, $\Delta = -2.28$; “a photo of a male scientist”, $\Delta = -2.22$; “she is a call center agent”, $\Delta = -2.06$), where male subjects receive disproportionately greater face attention. For scientist prompts, this pattern is consistent with gender-role congruence operating in the opposite direction: when the scientist label is paired with a male subject, the model anchors on facial features as the primary identity-confirming signal, with the occupational prop (lab coat) providing insufficient additional disambiguation. For call center agent prompts, the negative gap persists across both female- and male-gendered variants (“she is a call center agent”, $\Delta = -2.06$; “he is a call center agent”, $\Delta = -1.67$), suggesting that the headset prop is insufficiently distinctive to modulate the gender effect, and that male subjects’ faces carry higher diagnostic weight for this occupationally ambiguous role.

A theoretically important finding is that several male-gendered prompts (“he is a chef”, $\Delta = +0.26$; “he is a barista”, $\Delta = +0.15$) nevertheless produce positive gaps, indicating that female subjects receive greater face attention even under prompts with male gender markers. To rigorously assess whether the elevated FoR of female subjects

Race-Explicit Prompt Comparison — Black Man | White Man

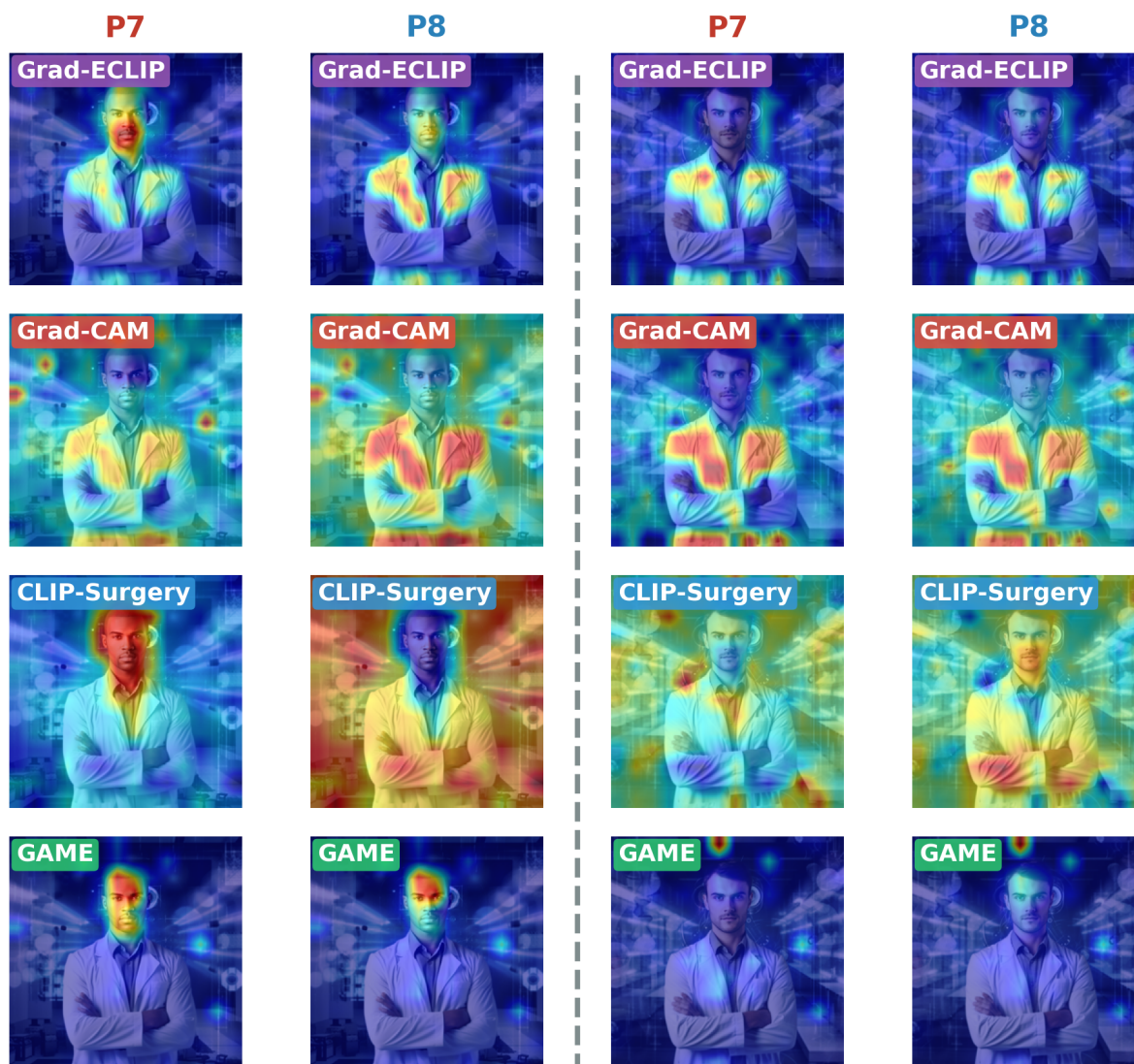


Figure 7: Saliency maps for Race-explicit prompts P7 (“a photo of the black scientist”) and P8 (“a photo of the white scientist”) across four attribution methods, for Black male (left) and White male (right) subjects wearing a lab coat. Each row corresponds to one attribution method (Grad-ECLIP, Grad-CAM, CLIP-Surgery, GAME) each column corresponds to one prompt–subject combination.

**Race Gap in FoR (ratio_{ses})
Across Race-Explicit Prompts and Methods**

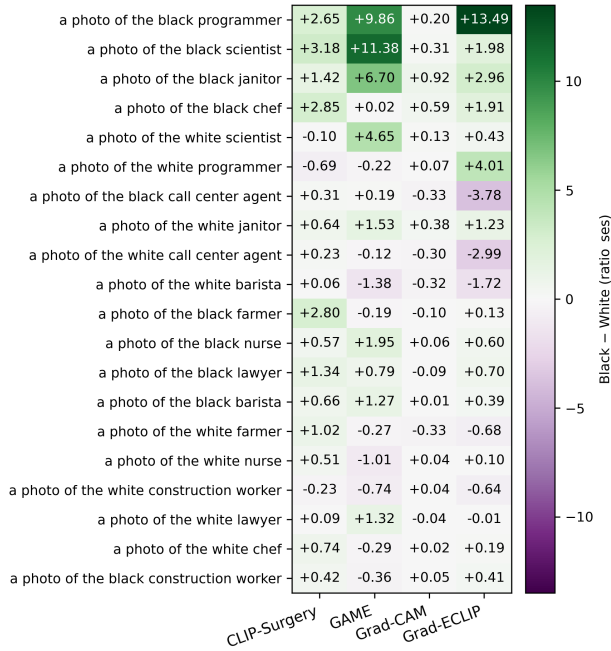


Figure 8: Race Gap heatmap in FoR (or *ratio_{ses}*) across race-explicit prompts for all methods.

under male-gendered prompts reflects a systematic pattern rather than isolated observations, further statistical investigation is warranted.

Comparison Between Explanation Methods (RQ3)

We observe that GAME produces systematically higher FoR values with greater variance, while Grad-CAM consistently falls below the neutral threshold. Rather than treating these as contradictory findings, we interpret them as reflecting the different computational substrates each method accesses. Grad-CAM’s spatial averaging over convolutional feature maps is known to favor large homogeneous regions, which in our setting corresponds to clothing and background. GAME’s attention-weighted ensemble is more sensitive to fine-grained spatial weighting but also more susceptible to head-specific noise. We therefore treat cross-method consistency as our primary criterion for a robust finding, rather than the absolute magnitude of any single method.

CLIP-Surgery and Grad-ECLIP are specifically designed to handle CLIP’s architectural quirks, so they tend to produce more semantically coherent maps for contrastive text-image matching tasks.

Limitations

Several limitations should be considered when interpreting these findings. The PAIRS dataset covers 20 occupation scenarios with four demographic subjects each, generated rather than real-world images, and limited to two racial

**Mean Gender Gap
in Face-Occupation Attention Ratio per Prompt**

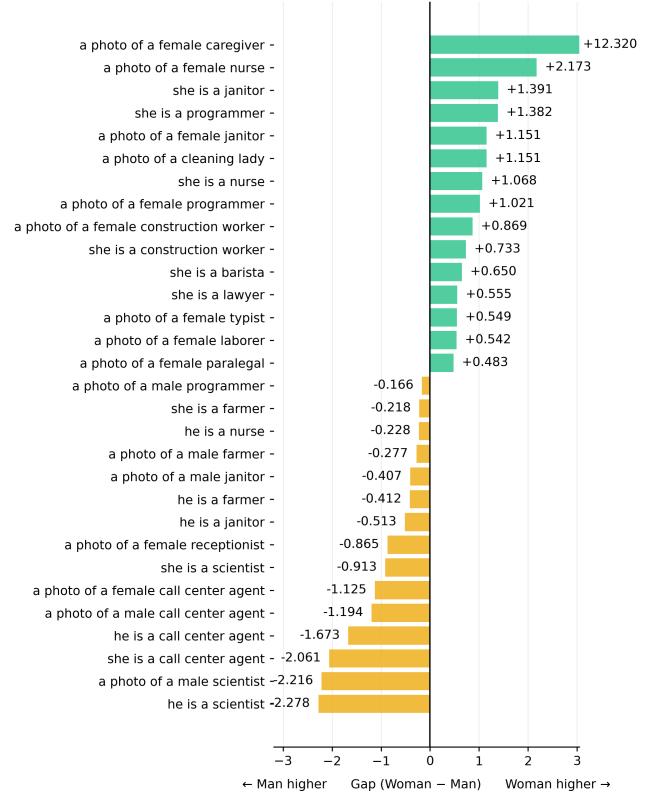


Figure 9: The figure displays the mean gender gap in the FoR for the 60 prompts with the largest absolute gaps (top 15 per direction), where positive values (*green, rightward*) indicate that female subjects receive relatively more face attention than male subjects under the same prompt, and negative values (*amber, leftward*) indicate the reverse.

categories, excluding other racial groups, age variation, and disability status. Conclusions should therefore be treated as preliminary and subject to validation on larger, more diverse benchmarks. Experiments are conducted exclusively on CLIP ViT-B/16, and findings may not generalize to other VLM architectures. Each attribution method carries known limitations; no single method is universally reliable, and the directional inconsistencies observed across methods reflect this inherent variability rather than contradicting the overall pattern of results. A further limitation concerns the absence of human attention validation. The attribution maps analyzed here reflect model-internal computational processes rather than human perceptual attention, and no ground-truth data on how human observers attend to the same stimuli was collected. Future work should incorporate human attention benchmarks, such as eye-tracking data or annotator saliency studies, to establish whether model attribution patterns diverge systematically from human perceptual priorities across demographic conditions.

Finally, the metric framework introduced here captures region-level attention density but does not address all di-

mensions of spatial attribution behavior. Further metrics and evaluation protocols remain an important direction for future work.

Conclusion

In this study we examined identity-conditioned attentional asymmetries in CLIP through a systematic analysis of post-hoc visual attribution outputs across occupational prompt conditions varying in racial and gender identity, with particular attention to how these asymmetries depend on prompt formulation. To support this analysis, a novel metric framework was introduced comprising the **Saliency Enrichment Score** (SES) and the **Face–Occupation Ratio** (FoR) to quantify where model attention is directed and how stably that direction is maintained across prompt formulations. A structured evaluation pipeline was applied across four attribution methods (Grad-ECLIP, Grad-CAM, CLIP-Surgery, and GAME) enabling cross-method assessment of whether observed patterns reflect stable model properties or methodological artifacts. No single method dominates across all evaluation criteria; each captures a distinct facet of the model’s attribution behavior, and conclusions drawn from any one method alone would be substantively incomplete.

The results show that the model does not attend to occupational features uniformly across demographic groups, and identity tokens modulate where and how stably the model attends to the scene. This modulation is most clearly patterned along racial and gender lines under explicit identity prompting, while subject-level differences under neutral occupational prompts are directionally consistent but not statistically separable at the aggregate level. This entanglement between identity and occupational representation carries direct practical implications for downstream applications of vision-language models in high-stakes professional contexts: including hiring support systems, credential verification pipelines, and clinical image tools.

This study was motivated by three research questions. **RQ1** is answered with qualification: saliency differences across demographic groups are clearest under race-explicit prompting (P7 vs. P8), where the strongest race-conditioned differences in FoR are observed, alongside cross-method variability. At the subject level, under identical occupational prompts, Black subject conditions show directionally higher faithfulness scores and greater FoR volatility than White subject conditions, but bootstrap confidence intervals for mean FoR overlap across all four subject conditions (Figure 6), so this difference is not statistically separable at the aggregate level. **RQ1** is therefore supported for prompt-explicit identity conditioning, while the subject-level effect remains a directional trend requiring further statistical power to confirm. **RQ2** is likewise confirmed: both race- and gender-explicit prompt modifications amplify observed differences in saliency distribution relative to identity-anchored conditions. And finally, **RQ3** reveals partial cross-method agreement: Grad-ECLIP, CLIP-Surgery, and GAME produce consistent directional signals, while Grad-CAM constitutes an exception, and methods diverge substantially in magnitude; reinforcing the view that cross-method convergence should be treated as an evidential standard rather than an optional

check in attribution-based bias auditing. The metric framework and evaluation pipeline establish an empirical baseline for future attribution-based auditing of vision-language models. Future investigations should extend this framework to real-world datasets, including photographs of individuals spanning additional demographic axes such as age, disability, and intersectional race combinations. This represents a natural and necessary direction for comprehensive model auditing.

Appendix: Prompt Generation Protocol

Benchmark prompts were generated using GPT-4o (gpt-4o-2024-05-13) under a fixed meta-prompt, reproduced in full in the supplementary material, with temperature set to 0.1 to minimize output variability across runs. The meta-prompt specified an 11-position prompt structure per subject: an identity-anchored opening position, seven shared occupation/gender/race-variant positions, and three scenario-specific role-variant positions. All candidate prompts were reviewed by the authors against three inclusion criteria: **lexical naturalness** (the prompt reads as a plausible natural language description), **affective neutrality** (no evaluative or connotative language beyond occupational and identity descriptors), and **structural parallelism** (the syntactic position of the identity token is consistent across demographic variants). Prompts failing any criterion were revised or discarded by author consensus. The final benchmark comprises 11 prompts per subject across 20 occupational scenarios and 4 demographic conditions, yielding 880 image-prompt pairs in total.

References

- Agarwal, S.; Krueger, G.; Clark, J.; Radford, A.; Kim, J. W.; and Brundage, M. 2021. Evaluating CLIP: Towards Characterization of Broader Capabilities and Downstream Implications. *ArXiv*, abs/2108.02818.
- Alayrac, J.-B.; Donahue, J.; Luc, P.; Miech, A.; Barr, I.; Hasson, Y.; Lenc, K.; Mensch, A.; Millican, K.; Reynolds, M.; Ring, R.; Rutherford, E.; Cabi, S.; Han, T.; Gong, Z.; Samangooei, S.; Monteiro, M.; Menick, J.; Borgeaud, S.; Brock, A.; Nematzadeh, A.; Sharifzadeh, S.; Binkowski, M.; Barreira, R.; Vinyals, O.; Zisserman, A.; and Simonyan, K. 2022. Flamingo: a Visual Language Model for Few-Shot Learning. In *Advances in Neural Information Processing Systems*, volume 35. NeurIPS 2022.
- Caliskan, A.; Bryson, J. J.; and Narayanan, A. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334): 183–186.
- Carion, N.; Gustafson, L.; Hu, Y.-T.; et al. 2026. SAM 3: Segment Anything with Concepts. In *The Fourteenth International Conference on Learning Representations*.
- Chefer, H.; Gur, S.; and Wolf, L. 2021. Generic Attention-model Explainability for Interpreting Bi-Modal and Encoder-Decoder Transformers. In *2021 IEEE/CVF International Conference on Computer Vision (ICCV)*, 387–396.

- Danish, S.; Sadeghi-Niaraki, A.; Khan, S. U.; Dang, L. M.; Tightiz, L.; and Moon, H. 2026. A comprehensive survey of Vision–Language Models: Pretrained models, fine-tuning, prompt engineering, adapters, and benchmark datasets. *Information Fusion*, 126: 103623.
- Dehouche, N. 2021. Implicit Stereotypes in Pre-Trained Classifiers. *IEEE Access*, 9: 167936–167947.
- Dwyer, B.; Nelson, J.; and Hansen, e. a., T. 2026. Roboflow (Version 1.0) [Software]. Available from <https://roboflow.com>. Computer vision.
- Fine, A.; Berthelot, E. R.; and Marsh, S. 2025. Public Perceptions of Judges’ Use of AI Tools in Courtroom Decision-Making: An Examination of Legitimacy, Fairness, Trust, and Procedural Justice. *Behavioral Sciences*, 15(4): 476.
- Fraser, K.; and Kiritchenko, S. 2024. Examining Gender and Racial Bias in Large Vision–Language Models Using a Novel Dataset of Parallel Images. In Graham, Y.; and Purver, M., eds., *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, 690–713. St. Julian’s, Malta: Association for Computational Linguistics.
- Hamidieh, K.; Zhang, H.; Gerych, W.; Hartvigsen, T.; and Ghassemi, M. 2024. Identifying Implicit Social Biases in Vision-Language Models. In *Proceedings of the Seventh AAAI/ACM Conference on AI, Ethics, and Society*, AIES-24, 547–561.
- Lee, N.; Bang, Y.; Lovenia, H.; Cahyawijaya, S.; Dai, W.; and Fung, P. 2023. Survey of Social Bias in Vision-Language Models. [arXiv:2309.14381](https://arxiv.org/abs/2309.14381).
- Li, J.; Li, D.; Savarese, S.; and Hoi, S. 2023. BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *ICML ’23*, 19730–19742.
- Li, Y.; Wang, H.; Duan, Y.; Xu, H.; and Li, X. 2022. Exploring visual interpretability for contrastive language-image pre-training. *arXiv preprint arXiv:2209.07046*.
- Li, Y.; Wang, H.; Duan, Y.; Zhang, J.; and Li, X. 2025. A closer look at the explainability of Contrastive language-image pre-training. *Pattern Recognition*, 162: 111409.
- Liu, H.; Li, C.; Wu, Q.; and Lee, Y. J. 2023. Visual Instruction Tuning. In *Advances in Neural Information Processing Systems*, volume 36. NeurIPS 2023 Oral.
- Marani, B.-E.; Hanini, M.; Malayarukil, N.; Christodoulidis, S.; Vakalopoulou, M.; and Ferrante, E. 2024. ViG-Bias: Visually Grounded Bias Discovery and Mitigation. In *European Conference on Computer Vision*. Springer Nature Switzerland.
- Narayanan, A.; Khazaie, V. R.; and Raza, S. 2025. Bias in the Picture: Benchmarking VLMs with Social-Cue News Images and LLM-as-Judge Assessment. [arXiv:2509.19659](https://arxiv.org/abs/2509.19659).
- Nozza, D.; Bianchi, F.; and Hovy, D. 2021. HONEST: Measuring Hurtful Sentence Completion in Language Models. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2398–2406. Online: Association for Computational Linguistics.
- OpenAI. 2024. GPT-4o System Card. *arXiv preprint arXiv:2410.21276*.
- Petsiuk, V.; Das, A.; and Saenko, K. 2018. RISE: Randomized Input Sampling for Explanation of Black-box Models. In *British Machine Vision Conference (BMVC)*.
- Radford, A.; Kim, J. W.; Hallacy, C.; Ramesh, A.; Goh, G.; Agarwal, S.; Sastry, G.; Askell, A.; Mishkin, P.; Clark, J.; Krueger, G.; and Sutskever, I. 2021. Learning Transferable Visual Models From Natural Language Supervision. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *PMLR*. ICML 2021 Oral.
- Ramachandranpillai, R.; Baeza-Yates, R.; and Heintz, F. 2025. FairXAI - A Taxonomy and Framework for Fairness and Explainability Synergy in Machine Learning. *IEEE Transactions on Neural Networks and Learning Systems*, 36(6): 9819–9836.
- Ruggeri, G.; Nozza, D.; et al. 2023. A multi-dimensional study on bias in vision-language models. In *Findings of the Association for Computational Linguistics: ACL 2023*. Association for Computational Linguistics.
- Schellmann, H. 2024. The Algorithm: How AI Decides Who Gets Hired. Hachette Book Group.
- Selvaraju, R. R.; Cogswell, M.; Das, A.; Vedantam, R.; Parikh, D.; and Batra, D. 2017. Grad-cam: Visual explanations from deep networks via gradient-based localization. In *Proceedings of the IEEE international conference on computer vision*, 618–626.
- Vo, A.; Nguyen, K.-N.; Taesiri, M. R.; Dang, V. T.; Nguyen, A. T.; and Kim, D. 2025. Vision Language Models are Biased. [arXiv:2505.23941](https://arxiv.org/abs/2505.23941).
- Wang, B.; Chen, W.; Pei, H.; Xie, C.; Kang, M.; Zhang, C.; Xu, C.; Xiong, Z.; Dutta, R.; Schaeffer, R.; Truong, S. T.; Arora, S.; Mazeika, M.; Hendrycks, D.; Lin, Z.; Cheng, Y.; Koyejo, S.; Song, D.; and Li, B. 2024. DecodingTrust: A Comprehensive Assessment of Trustworthiness in GPT Models. [arXiv:2306.11698](https://arxiv.org/abs/2306.11698).
- Weng, Z.; Gao, Z.; Andrews, J.; and Zhao, J. 2024. Images Speak Louder than Words: Understanding and Mitigating Bias in Vision-Language Model from a Causal Mediation Perspective. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, 15669–15680. Miami, Florida, USA: Association for Computational Linguistics.
- Wolfe, R.; Banaji, M. R.; and Caliskan, A. 2022. Evidence for Hypodescent in Visual Semantic AI. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT ’22, 1293–1304. New York, NY, USA: Association for Computing Machinery. ISBN 9781450393522.
- Yang, Y.; et al. 2025. Demographic Bias of Expert-Level Vision-Language Foundation Models in Medical Imaging. *Science Advances*, 11: eadq0305.
- Zhao, C.; Wang, K.; Hsiao, J. H.; and Chan, A. B. 2025. Grad-ECLIP: Gradient-based Visual and Textual Explanations for CLIP. [arXiv:2502.18816](https://arxiv.org/abs/2502.18816).

Zhao, C.; Wang, K.; Zeng, X.; Zhao, R.; and Chan, A. B. 2024. Gradient-based visual explanation for transformer-based CLIP. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Zhong, Y.; Yang, J.; Zhang, P.; Li, C.; Codella, N.; Li, L. H.; Zhou, L.; Dai, X.; Yuan, L.; Li, Y.; and Gao, J. 2022. RegionCLIP: Region-Based Language-Image Pretraining. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, CVPR 2022, 16793–16803.