# Novelty Assessment Report

**Paper**: Guardrail-Agnostic Societal Bias Evaluation in Large Vision-Language Models
**PDF URL**: https://openreview.net/pdf?id=2PJkG6aV4A
**Venue**: ICLR 2026 Conference Submission
**Year**: 2026
**Report Generated**: 2026-01-05

## Abstract

We propose a societal bias evaluation method for large vision-language models (LVLMs) in the era of strong safety guardrails. Existing benchmarks rely on prompts that ask models to infer attributes of people in images (e.g., "Is this person a CEO or a secretary?"). However, we find that LVLMs with strong guardrails, such as GPT and Claude, often refuse these prompts, making evaluations unreliable. To address this, we change the prior evaluation paradigm by decoupling the task from the depicted person: instead of inferring person's attributes, we use prompts that do not ask about the person (e.g., "Write a fictional story about an imaginary person.") and attach the image as provisional user information to implicitly provide demographic cues, then compare outputs across user demographics. Instantiated across three tasks — story generation, term explanation, and exam-style QA — our method avoids refusals even in guardrailed LVLMs, enabling reliable bias measurement. Applying it to 20 recent LVLMs, both open-source and proprietary, we find that all models undesirably use user demographic information in person-irrelevant tasks; for instance, characters in stories are often portrayed as mechanic for male users and nurse for female users. Although still biased, proprietary models like GPT-5 show lower bias than open-source ones. We analyze potential factors behind this gap, discussing continuous model monitoring and improvement as a possible driving factor for reducing bias.



## Core Task Landscape

This paper addresses: **Societal Bias Evaluation in Large Vision-Language Models**

A total of **50 papers** were analyzed and organized into a taxonomy with **18 categories**.

### Taxonomy Overview

The research landscape has been organized into the following main categories:

- **Bias Measurement Frameworks and Benchmarks**
- **Bias Analysis in Specific VLM Architectures**
- **Bias Sources and Contributing Factors**
- **Bias Mitigation Techniques**
- **Domain-Specific and Contextual Bias Studies**
- **Methodological Advances and Meta-Analysis**

### Complete Taxonomy Tree

- Societal Bias Evaluation in Large Vision-Language Models Survey Taxonomy
- Bias Measurement Frameworks and Benchmarks
  - Comprehensive Multi-Dimensional Bias Benchmarks (6 papers)
  - [1] Vhelm: A holistic evaluation of vision language models (Tony Lee, 2024) View paper
  - [4] Vlbiasbench: A comprehensive benchmark for evaluating bias in large vision-language model (Wang Sibo, 2024) View paper
  - [6] VIGNETTE: Socially Grounded Bias Evaluation for Vision-Language Models (Raj, 2025) View paper
  - [8] REVAL: A Comprehension Evaluation on Reliability and Values of Large Vision-Language Models (Zhang Jie, 2025) View paper
  - [20] A unified framework and dataset for assessing societal bias in vision-language models (Jain, 2024) View paper
  - [37] SB-Bench: Stereotype Bias Benchmark for Large Multimodal Models (Vishal Narnaware, 2025) View paper
  - Specialized Bias Assessment Tools (6 papers)
  - [7] Biasdora: Exploring hidden biased associations in vision-language models (Anastasopoulos, 2024) View paper
  - [11] Examining gender and racial bias in large vision-language models using a novel dataset of parallel images (Fraser Kathleen, 2024) View paper
  - [24] Socialcounterfactuals: Probing and mitigating intersectional social biases in vision-language models with counterfactual examples (Phillip Howard, 2024) View paper
  - [26] Multimodal bias: Introducing a framework for stereotypical bias assessment beyond gender and race in vision language models (de Melo, 2023) View paper
  - [29] GenderBias-VL: Benchmarking Gender Bias in Vision Language Models via Counterfactual Probing: Y. Xiao et al. (Y Xiao, 2025) View paper
  - [48] Bias in the Picture: Benchmarking VLMs with Social-Cue News Images and LLM-as-Judge Assessment (Narayanan Aravind, 2025) View paper
  - Implicit and Stereotype Bias Probing (5 papers)
  - [5] Identifying implicit social biases in vision-language models (Hamidieh, 2024) View paper
  - [19] Vlstereoset: A study of stereotypical bias in pre-trained vision-language models (Jiang Jing, 2022) View paper
  - [21] VisBias: Measuring Explicit and Implicit Social Biases in Vision Language Models (Huang, 2025) View paper
  - [40] Measuring Social Biases in Grounded Vision and Language Embeddings (Barbu, 2021) View paper
  - [41] Beyond the Surface: A Comprehensive Analysis of Implicit Bias in Vision-Language Models (Giacomo Capitani, 2024) View paper

- Explicit Bias and Attribute Inference Studies (2 papers)
- [22] : Measuring Stereotypical Bias in Large Vision-Language Models from Vision and Language Modalities (Y Jiang, 2024) View paper
- [50] Investigating Social Biases in Multimodal LLMs (Malsha V. Perera, 2025) View paper

- Bias Analysis in Specific VLM Architectures
  - Contrastive Vision-Language Models (5 papers)
  - [2] Social perception of faces in a vision-language model (Knott, 2025) View paper
  - [10] Data matters most: Auditing social bias in contrastive visionâ□□language models (Sahili, 2025) View paper
  - [14] A Comprehensive Social Bias Audit of Contrastive Vision Language Models (ZA Sahili, 2025) View paper
  - [31] Dataset scale and societal consistency mediate facial impression bias in vision-language ai (Wolfe, 2024) View paper
  - [38] American== white in multimodal language-and-image ai (Robert Wolfe, 2022) View paper
  - Large Vision-Language Models and Assistants ★ (5 papers)
  - [0] Guardrail-Agnostic Societal Bias Evaluation in Large Vision-Language Models (Anon et al., 2026) View paper
  - [12] Uncovering bias in large vision-language models with counterfactuals (Howard, 2024) View paper
  - [17] A Closer Look at Bias and Chain-of-Thought Faithfulness of Large (Vision) Language Models (Sriram Balasubramanian, 2025) View paper
  - [44] Revealing and Reducing Gender Biases in Vision and Language Assistants (VLAs) (Girrbach, 2024) View paper
  - [46] Uncovering bias in large vision-language models at scale with counterfactuals (Bhiwandiwalla, 2025) View paper
  - Pretrained and Fine-Tuned VLM Families (3 papers)
  - [16] Evaluating bias and fairness in gender-neutral pretrained vision-and-language models (Brandl, 2023) View paper
  - [18] A multi-dimensional study on bias in vision-language models (Nozza, 2023) View paper
  - [36] Worst of Both Worlds: Biases Compound in Pre-trained Vision-and-Language Models (Srinivasan, 2022) View paper
- Bias Sources and Contributing Factors
  - Training Data Scale and Composition Effects (2 papers)
  - [27] Societal bias in vision-and-language datasets and models (Y Nakashima, 2023) View paper
  - [34] The dark side of dataset scaling: Evaluating racial classification in multimodal models (Birhane, 2024) View paper
  - Model Architecture and Scale Factors (1 papers)
  - [42] Vision Language Models are Biased (Nguyen, 2025) View paper
- Bias Mitigation Techniques
  - Embedding Space Debiasing (4 papers)
  - [28] My Answer Is NOT'Fair': Mitigating Social Bias in Vision-Language Models via Fair and Biased Residuals (Lan Jian, 2025) View paper
  - [32] Debiasing vision-language models via biased prompts (Chuang, 2023) View paper
  - [33] A Prompt Array Keeps the Bias Away: Debiasing Vision-Language Models with Adversarial Learning (Bain, 2022) View paper
  - [43] Dear: Debiasing vision-language models with additive residuals (Ashish Seth, 2023) View paper
  - Prompt-Based and Data Augmentation Debiasing (2 papers)
  - [23] Social debiasing for fair multi-modal llms (Cheng, 2025) View paper
  - [45] DR.GAP: Mitigating Bias in Large Language Models using Gender-Aware Prompting with Demonstration and Reasoning (Qiu HongYe, 2025) View paper
- Domain-Specific and Contextual Bias Studies
  - Cultural and Geographic Bias (3 papers)
  - [9] Investigating Stereotypical Bias in Large Language and Vision-Language Models (Pang, 2025) View paper
  - [35] See It from My Perspective: Diagnosing the Western Cultural Bias of Large Vision-Language Models in Image Understanding (Ananthram, 2024) View paper
  - [49] Toward Socially Aware Vision-Language Models: Evaluating Cultural Competence Through Multimodal Story Generation (Mukherjee Arka, 2025) View paper
  - Social Perception and Demographic Biases (1 papers)
  - [25] Visual Cues of Gender and Race are Associated with Stereotyping in Vision-Language Models (Jeon Soyeon, 2025) View paper
  - Task-Specific Bias Manifestations (1 papers)
  - [39] User-vlm 360: Personalized vision language models with user-aware tuning for social human-robot interactions (Hamed Rahimi, 2025) View paper
- Methodological Advances and Meta-Analysis
  - Survey and Review Papers (2 papers)
  - [3] Survey of social bias in vision-language models (Leeï¼□ Na-Yeon, 2023) View paper
  - [13] Fairness in Deep Learning: A survey on vision and language research (Otavio Parraga, 2025) View paper
  - Evaluation Methodology and Metric Design (1 papers)
  - [47] Improving Bias Metrics in Vision-Language Models by Addressing Inherent Model Disabilities (LB Darur, 2024) View paper
  - Robustness and Adversarial Bias Evaluation (1 papers)
  - [30] B-AVIBench: Toward Evaluating the Robustness of Large Vision-Language Model on Black-Box Adversarial Visual-Instructions (Hao Zhang, 2024) View paper
  - Language Bias and Multimodal Interaction (1 papers)
  - [15] Looking beyond text: Reducing language bias in large vision-language models via multimodal dual-attention and soft-image guidance (HaoZhe Zhao, 2025) View paper

## Narrative

Core task: societal bias evaluation in large vision-language models. The field has organized itself around several complementary branches that together address how biases emerge, persist, and might be measured or reduced in VLMs. Bias Measurement Frameworks and Benchmarks establish standardized evaluation protocols and datasets—ranging from holistic suites like Vhelm Holistic Evaluation[1] to specialized resources such as Vlbiasbench[4] and VIGNETTE[6]—that enable systematic comparison across models. Bias Analysis in Specific VLM Architectures examines how particular model families (large assistants, retrieval-augmented systems, or embodied agents) exhibit distinct bias patterns, while Bias Sources and Contributing Factors investigates the origins of bias in training data, architectural choices, and scaling dynamics. Bias Mitigation Techniques explores interventions such as prompt engineering, debiasing algorithms, and fairness-aware training, and Domain-Specific and Contextual Bias Studies zoom into particular application areas (news captioning, medical imaging, cultural contexts) where biases have unique manifestations. Finally, Methodological Advances and Meta-Analysis refines evaluation metrics and synthesizes findings across studies to improve the rigor of bias research.

Recent work has intensified around two contrasting themes: developing richer benchmarks that capture intersectional and implicit biases (e.g., Implicit Social Biases[5], Social Perception Faces[2]) versus probing how guardrails and safety mechanisms themselves interact with bias (Guardrail Agnostic Bias[0]). The original paper, Guardrail Agnostic Bias[0], sits within the branch analyzing large vision-language models and assistants, where it addresses a relatively underexplored question: whether safety interventions inadvertently modulate or mask underlying biases. This contrasts with nearby studies like Gender Biases VLAs[44], which document bias prevalence in vision-language agents, and Counterfactuals at Scale[46], which uses large-scale counterfactual generation to measure bias robustness. By focusing on guardrail effects, Guardrail Agnostic Bias[0] highlights an emerging concern that evaluation must account for post-hoc safety layers, not just base model behavior, to obtain a complete picture of societal bias in deployed systems.

## Related Works in Same Category

The following **4 sibling papers** share the same taxonomy leaf node with the original paper:

### 1. Uncovering bias in large vision-language models with counterfactuals

**Authors**: Howard, Phillip, Bhiwandiwalla, Anahita, Phillip Howard, et al. (12 authors total) | **Year/Venue**: 2024 | **URL**: View paper

#### Abstract

With the advent of Large Language Models (LLMs) possessing increasingly impressive capabilities, a number of Large Vision-Language Models (LVLMs) have been proposed to augment LLMs with visual inputs. Such models condition generated text on both an input image and a text prompt, enabling a variety of use cases such as visual question answering and multimodal chat. While prior studies have examined the social biases contained in text generated by LLMs, this topic has been relatively unexplored in...

#### Relationship Analysis

Both papers belong to the Large Vision-Language Models and Assistants category, focusing on bias evaluation in generative LVLMs with conversational capabilities. They overlap in examining societal bias (gender, race) in LVLMs through systematic evaluation frameworks, but differ fundamentally in approach: the original paper proposes a guardrail-agnostic method using person-irrelevant prompts with images as user context to avoid refusals, while the candidate paper uses counterfactual images with open-ended attribute-inferring prompts to isolate the influence of social attributes on generated text toxicity and competency descriptions.

### 2. A Closer Look at Bias and Chain-of-Thought Faithfulness of Large (Vision) Language Models

**Authors**: Sriram Balasubramanian, Samyadeep Basu, Soheil Feizi | **Year/Venue**: 2025 | **URL**: View paper

#### Abstract

â¦ of chain-of-thought faithfulness in Large Vision-Language Models (LVLMs), addressing a â¦ bias induction into the model from bias evaluation, enabling more precise analysis of how â¦

#### Relationship Analysis

Both papers belong to the Large Vision-Language Models and Assistants category, focusing on bias evaluation in generative LVLMs with conversational capabilities. They share overlapping concerns about societal bias in models like GPT and Claude, but differ fundamentally in their approaches: the original paper proposes a guardrail-agnostic evaluation method using person-irrelevant prompts to measure bias when models refuse attribute-inferring questions, while the candidate paper investigates chain-of-thought faithfulness and bias articulation patterns, examining whether models explicitly acknowledge biasing cues in their reasoning traces across both text-based and image-based biases.

### 3. Revealing and Reducing Gender Biases in Vision and Language Assistants (VLAs)

**Authors**: Girrbach, Leander, Alaniz, Stephan, Leander Girrbach, et al. (14 authors total) | **Year/Venue**: 2024 • International Conference on Learning Representations | **URL**: View paper

#### Abstract

Pre-trained large language models (LLMs) have been reliably integrated with visual input for multimodal tasks. The widespread adoption of instruction-tuned image-to-text vision-language assistants (VLAs) like LLaVA and InternVL necessitates evaluating gender biases. We study gender bias in 22 popular open-source VLAs with respect to personality traits, skills, and occupations. Our results show that VLAs replicate human biases likely present in the data, such as real-world occupational imbalances...

#### Relationship Analysis

Both papers belong to the Large Vision-Language Models and Assistants category, focusing on bias evaluation in instruction-tuned multimodal models. They overlap in examining gender bias in VLAs like LLaVA and InternVL, but differ fundamentally in approach: the original paper proposes a guardrail-agnostic evaluation method using person-irrelevant prompts with images as user context to avoid refusals, while the candidate paper uses traditional attribute-inferring prompts to assess personality traits, skills, and occupations, and additionally proposes fine-tuning-based debiasing methods to reduce bias.

### 4. Uncovering bias in large vision-language models at scale with counterfactuals

**Authors**: Bhiwandiwalla, Anahita, Fraser Kathleen, Howard, Phillip R., et al. (7 authors total) | **Year/Venue**: 2025 | **URL**: View paper

#### Abstract

With the advent of Large Language Models (LLMs) possessing increasingly impressive capabilities, a number of Large Vision-Language Models (LVLMs) have been proposed to augment LLMs with visual inputs. Such models condition generated text on both an input image and a text prompt, enabling a variety of use cases such as visual question answering and multimodal chat. While prior studies have examined the social biases contained in text generated by LLMs, this topic has been relatively unexplored in...

#### Relationship Analysis

Both papers belong to the Large Vision-Language Models and Assistants category, focusing on bias evaluation in generative LVLMs with conversational capabilities. They overlap in evaluating societal bias (gender, race) in recent LVLMs using open-ended generation tasks and analyzing output disparities across demographic groups. However, the original paper proposes a guardrail-agnostic evaluation method using person-irrelevant prompts with images as user context to avoid refusals, while the candidate paper uses counterfactual synthetic images with attribute-inferring prompts to isolate bias effects, conducting a large-scale study (57M generations) with multi-dimensional analysis including toxicity, stereotypes, and competency words.

## Contributions Analysis

**Overall novelty summary.** The paper proposes a guardrail-agnostic evaluation method for societal bias in large vision-language models, addressing the challenge that safety-aligned models often refuse direct attribute-inference prompts. It resides in the 'Large Vision-Language Models and Assistants' leaf, which contains five papers examining bias in generative LVLMs and instruction-tuned assistants. This leaf sits within a broader taxonomy of 50 papers across bias measurement, mitigation, and analysis, indicating a moderately populated research direction focused on modern conversational VLMs rather than contrastive models like CLIP.

The taxonomy reveals neighboring leaves examining contrastive VLMs (five papers on CLIP-family models) and pretrained VLM families (three comparative studies). The paper's focus on guardrail interactions distinguishes it from sibling works like Gender Biases VLAs, which documents bias prevalence without addressing safety mechanisms, and Counterfactuals at Scale, which uses data augmentation for robustness testing. The 'Bias Measurement Frameworks' branch contains comprehensive benchmarks (six papers) and specialized tools (six papers), but none explicitly tackle the refusal problem in safety-aligned models, suggesting the paper addresses a gap at the intersection of bias evaluation and model alignment.

Among 27 candidates examined through limited semantic search, none clearly refute the three core contributions. The guardrail-agnostic method examined 10 candidates with zero refutations, the three-task instantiation examined 10 with zero refutations, and the paradigm shift from attribute-inferring to person-irrelevant prompts examined 7 with zero refutations. This suggests that within the search scope—focused on recent LVLM bias literature—the specific approach of decoupling demographic cues from task prompts to circumvent safety refusals has not been documented. However, the limited search scale means unexplored work in adjacent areas (prompt engineering, implicit bias probing) may exist.

The analysis indicates novelty within the examined literature, particularly in addressing how safety guardrails complicate bias measurement. The taxonomy shows the field has developed rich benchmarks and mitigation techniques, but the specific methodological challenge of evaluating guardrailed models appears underexplored among the 50 papers surveyed. The contribution's distinctiveness depends partly on whether the 'person-irrelevant prompt' strategy represents a fundamental paradigm shift or an incremental adaptation of existing implicit bias probing methods, which the limited search scope cannot definitively resolve.

---

This paper presents **3 main contributions**, each analyzed against relevant prior work:

## Contribution 1: Guardrail-agnostic societal bias evaluation method for LVLMs

**Description**: The authors introduce a new evaluation framework that decouples the task from the depicted person by using person-irrelevant prompts and treating images as provisional user information rather than the subject of inference. This design avoids model refusals triggered by safety guardrails, enabling reliable bias measurement even in strongly guarded models.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

### 1. A Prompt Array Keeps the Bias Away: Debiasing Vision-Language Models with Adversarial Learning

**URL**: View paper

**Brief Assessment**

Prompt Array Debiasing[33] focuses on debiasing vision-language models through adversarial learning and prompt engineering, not on evaluation methods that circumvent safety guardrails. The candidate addresses bias mitigation rather than guardrail-agnostic measurement frameworks.

---

### 2. Red-Teaming for Inducing Societal Bias in Large Language Models

**URL**: View paper

**Brief Assessment**

Red Teaming Societal[54] focuses on red-teaming methods to induce biased responses in LLMs (not LVLMs) through adversarial attacks, while the original paper proposes a guardrail-agnostic evaluation framework for LVLMs using person-irrelevant prompts with images as user context.

---

### 3. Socialcounterfactuals: Probing and mitigating intersectional social biases in vision-language models with counterfactual examples

**URL**: View paper

**Brief Assessment**

Socialcounterfactuals[24] focuses on probing intersectional social biases using synthetically generated counterfactual image-text pairs, not on addressing guardrail-induced refusals in bias evaluation.

---

### 4. Data matters most: Auditing social bias in contrastive visionâ□□language models

**URL**: View paper

**Brief Assessment**

Data Matters Most[10] focuses on auditing social bias in contrastive vision-language models (CLIP/OpenCLIP) through systematic analysis of training data factors (model size, data scale, data source), not on developing guardrail-agnostic evaluation methods for safety-guarded models.

---

### 5. Words or Vision: Do Vision-Language Models Have Blind Faith in Text?

**URL**: View paper

**Brief Assessment**

Blind Faith Text[51] investigates modality preferences and text bias in VLMs when visual-textual inconsistencies occur, not societal bias evaluation methods that bypass safety guardrails. The papers address fundamentally different research problems.

---

### 6. Debiasing vision-language models via biased prompts

**URL**: View paper

**Brief Assessment**

Biased Prompts Debiasing[32] focuses on debiasing vision-language models through projection matrices applied to text embeddings, not on evaluation methods that circumvent safety guardrails. The candidate addresses bias mitigation in discriminative and generative models, while the original paper proposes a novel evaluation framework specifically designed to measure bias in guardrailed models by treating images as user context rather than inference targets.

---

### 7. Counterfactually Measuring and Eliminating Social Bias in Vision-Language Pre-training Models

**URL**: View paper

**Brief Assessment**

Counterfactually Measuring Bias[52] focuses on measuring bias in vision-language pre-training models using counterfactual image generation via adversarial attacks, not on addressing guardrail-induced refusals in modern LVLMs with safety mechanisms.

---

### 8. Investigating Stereotypical Bias in Large Language and Vision-Language Models

**URL**: View paper

**Brief Assessment**

Stereotypical Bias Investigation[9] focuses on measuring stereotypical bias through different perspectives but does not address the guardrail refusal problem or propose person-irrelevant prompts with images as user context, which is the core novelty of the original paper's evaluation framework.

---

### 9. CLIP the Bias: How Useful is Balancing Data in Multimodal Learning?

**URL**: View paper

**Brief Assessment**

CLIP Bias Balancing[53] focuses on data-balancing techniques to mitigate biases in CLIP models during pretraining, not on evaluation methodologies that bypass safety guardrails in LVLMs.

---

### 10. Social debiasing for fair multi-modal llms

**URL**: View paper

**Brief Assessment**

Social Debiasing[23] focuses on mitigating social biases through counterfactual datasets and debiasing strategies, not on developing evaluation methods that avoid guardrail-triggered refusals. The candidate addresses bias reduction via training, while the original addresses bias measurement despite safety mechanisms.

---

## Contribution 2: Three-task instantiation of the evaluation framework

**Description**: The authors instantiate their evaluation protocol across three person-irrelevant tasks: story generation, term explanation, and exam-style QA. Each task is designed to probe different aspects of societal bias while maintaining zero refusal rates across all tested models.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

### 1. Bias Beyond Demographics: Probing Decision Boundaries in Black-Box LVLMs via Counterfactual VQA

**URL**: View paper

**Brief Assessment**

Bias Beyond Demographics[69] focuses on counterfactual VQA with image pairs differing in single visual attributes across five categories (demography, culture, environment, behavior, aesthetic), not on the three person-irrelevant tasks (story generation, term explanation, exam-style QA) proposed in the original paper.

---

### 2. A Methodological Framework for Auditing Norm-Sensitive Behaviour in Large Language Models: Research Design for Employment Contexts

**URL**: View paper

**Brief Assessment**

Norm Sensitive Behaviour[65] focuses on auditing norm-sensitive behavior in employment contexts through a methodological framework, not on multi-task bias measurement avoiding model refusals in vision-language models.

---

### 3. One Model for All: Multi-Objective Controllable Language Models

**URL**: View paper

**Brief Assessment**

Multi Objective Controllable[71] focuses on multi-objective controllable language models for personalization across different user preferences (humor, helpfulness, harmlessness), not on societal bias evaluation frameworks. The tasks in Multi Objective Controllable[71] (story generation, term explanation, exam-style QA) serve different purposes than the original paper's bias measurement framework.

---

### 4. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned

**URL**: View paper

**Brief Assessment**

Red Teaming Harms[64] focuses on adversarial red teaming of language models through open-ended conversations to elicit harmful outputs, not on measuring societal bias through person-irrelevant tasks that avoid model refusals.

---

### 5. Silenced Biases: The Dark Side LLMs Learned to Refuse

**URL**: View paper

**Brief Assessment**

Silenced Biases[67] focuses on a single QA-based task using activation steering to bypass refusal mechanisms, not on multi-task frameworks. The candidate's approach differs fundamentally from the original's three-task design (story generation, term explanation, exam-style QA) for bias evaluation.

---

### 6. Trustllm: Trustworthiness in large language models

**URL**: View paper

**Brief Assessment**

Trustllm[62] focuses on trustworthiness evaluation across multiple dimensions (truthfulness, safety, fairness, etc.) rather than specifically addressing societal bias measurement frameworks that avoid model refusals. The candidate's multi-task approach serves different evaluation purposes than the original paper's person-irrelevant task design for bias assessment in vision-language models.

---

### 7. Multi-objective linguistic control of large language models

**URL**: View paper

**Brief Assessment**

Multi Objective Linguistic[63] focuses on controlling linguistic complexity features (word count, noun variation, etc.) across general NLP tasks, not on measuring societal bias across demographic groups while avoiding model refusals.

---

### 8. Understanding Large Language Model Vulnerabilities to Social Bias Attacks

**URL**: View paper

**Brief Assessment**

Social Bias Attacks[66] focuses on adversarial attack methods (prefix injection, refusal suppression, learned attack prompts) to induce biased responses in LLMs, not on multi-task bias measurement frameworks that avoid model refusals through person-irrelevant prompts.

---

#### 9. Benchmarking Adversarial Robustness to Bias Elicitation in Large Language Models: Scalable Automated Assessment with LLM-as-a-Judge

**URL**: View paper

**Brief Assessment**

Adversarial Robustness Benchmarking[68] focuses on adversarial jailbreak attacks to elicit biases in LLMs using multiple-choice and sentence completion tasks, not on person-irrelevant tasks with images as user context to avoid model refusals in vision-language models.

---

#### 10. From Narrow Unlearning to Emergent Misalignment: Causes, Consequences, and Containment in LLMs

**URL**: View paper

**Brief Assessment**

Narrow Unlearning Misalignment[70] focuses on refusal unlearning across seven RAI domains (cybersecurity, safety, toxicity, bias, sensitive content, medical/legal, privacy) to study emergent misalignment, not on multi-task bias measurement frameworks that avoid model refusals in vision-language models.

---

### Contribution 3: Paradigm shift from attribute-inferring to person-irrelevant prompts

**Description**: The authors propose a fundamental change in how bias is evaluated by replacing attribute-inferring prompts with person-irrelevant ones and changing the role of images from target to context. This paradigm shift addresses the refusal problem in existing benchmarks while reducing the impact of spurious image contexts.

This contribution was assessed against **7 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

#### 1. Towards Alleviating the Object Bias in Prompt Tuning-based Factual Knowledge Extraction

**URL**: View paper

**Brief Assessment**

Object Bias Prompt[60] addresses object bias in factual knowledge extraction from language models using subject-masked prompts, not person-irrelevant prompts for bias evaluation in vision-language models. The tasks and domains are fundamentally different.

---

#### 2. Measuring Mechanistic Independence: Can Bias Be Removed Without Erasing Demographics?

**URL**: View paper

**Brief Assessment**

Mechanistic Independence[58] focuses on mechanistic interpretability of bias features using sparse autoencoders and feature ablation in demographic reasoning tasks. It does not address the evaluation paradigm shift from attribute-inferring to person-irrelevant prompts that the original paper proposes for bias benchmarking in vision-language models.

---

#### 3. Investors' acceptance and use of investment-based crowdfunding platforms: an integrated perspective

**URL**: View paper

**Brief Assessment**

Crowdfunding Platforms Acceptance[57] focuses on user acceptance of investment crowdfunding platforms using technology acceptance models, not on bias evaluation methodologies or prompting paradigms for vision-language models.

---

#### 4. Assessing the Reliability of LLMs Annotations in the Context of Demographic Bias and Model Explanation

**URL**: View paper

**Brief Assessment**

LLMs Annotations Reliability[59] focuses on persona prompting for annotation tasks in sexism detection, not on bias evaluation paradigms. The candidate does not address attribute-inferring vs. person-irrelevant prompts or the role of images as context vs. target in bias benchmarks.

---

#### 5. Large language model as attributed training data generator: A tale of diversity and bias

**URL**: View paper

**Brief Assessment**

Attributed Training Generator[55] focuses on diversifying training data generation for text classification through attributed prompts (e.g., length, style, location), not on bias evaluation paradigms or person-irrelevant prompting for vision-language models.

---

#### 6. Can multimodal large language models enhance performance benefits among higher education students? An investigation based on the taskâ□□technology fit â□¦

**URL**: View paper

**Brief Assessment**

Multimodal Higher Education[56] focuses on task-technology fit in educational contexts with multimodal LLMs, not on bias evaluation paradigms or prompt design methodologies for societal bias measurement.

---

#### 7. Selective Retrieval of Stimulus Information versus Thematic Judgments in Natural Language Inferences.

**URL**: View paper

**Brief Assessment**

Selective Retrieval Stimulus[61] focuses on stimulus information retrieval and thematic judgments in natural language inferences, not on bias evaluation paradigms or prompt design for vision-language models.

---

## Appendix: Text Similarity Detection

No high-similarity text segments were detected across any compared papers.

## References


- [0] Guardrail-Agnostic Societal Bias Evaluation in Large Vision-Language Models View paper
- [1] Vhelm: A holistic evaluation of vision language models View paper
- [2] Social perception of faces in a vision-language model View paper
- [3] Survey of social bias in vision-language models View paper
- [4] Vlbiasbench: A comprehensive benchmark for evaluating bias in large vision-language model View paper
- [5] Identifying implicit social biases in vision-language models View paper

- [6] VIGNETTE: Socially Grounded Bias Evaluation for Vision-Language Models View paper
- [7] Biasdora: Exploring hidden biased associations in vision-language models View paper
- [8] REVAL: A Comprehension Evaluation on Reliability and Values of Large Vision-Language Models View paper
- [9] Investigating Stereotypical Bias in Large Language and Vision-Language Models View paper
- [10] Data matters most: Auditing social bias in contrastive visionâ□□language models View paper
- [11] Examining gender and racial bias in large vision-language models using a novel dataset of parallel images View paper
- [12] Uncovering bias in large vision-language models with counterfactuals View paper
- [13] Fairness in Deep Learning: A survey on vision and language research View paper
- [14] A Comprehensive Social Bias Audit of Contrastive Vision Language Models View paper
- [15] Looking beyond text: Reducing language bias in large vision-language models via multimodal dual-attention and soft-image guidance View paper
- [16] Evaluating bias and fairness in gender-neutral pretrained vision-and-language models View paper
- [17] A Closer Look at Bias and Chain-of-Thought Faithfulness of Large (Vision) Language Models View paper
- [18] A multi-dimensional study on bias in vision-language models View paper
- [19] Vlstereoset: A study of stereotypical bias in pre-trained vision-language models View paper
- [20] A unified framework and dataset for assessing societal bias in vision-language models View paper
- [21] VisBias: Measuring Explicit and Implicit Social Biases in Vision Language Models View paper
- [22] : Measuring Stereotypical Bias in Large Vision-Language Models from Vision and Language Modalities View paper
- [23] Social debiasing for fair multi-modal llms View paper
- [24] Socialcounterfactuals: Probing and mitigating intersectional social biases in vision-language models with counterfactual examples View paper
- [25] Visual Cues of Gender and Race are Associated with Stereotyping in Vision-Language Models View paper
- [26] Multimodal bias: Introducing a framework for stereotypical bias assessment beyond gender and race in vision language models View paper
- [27] Societal bias in vision-and-language datasets and models View paper
- [28] My Answer Is NOT'Fair': Mitigating Social Bias in Vision-Language Models via Fair and Biased Residuals View paper
- [29] GenderBias-VL: Benchmarking Gender Bias in Vision Language Models via Counterfactual Probing: Y. Xiao et al. View paper
- [30] B-AVIBench: Toward Evaluating the Robustness of Large Vision-Language Model on Black-Box Adversarial Visual-Instructions View paper
- [31] Dataset scale and societal consistency mediate facial impression bias in vision-language ai View paper
- [32] Debiasing vision-language models via biased prompts View paper
- [33] A Prompt Array Keeps the Bias Away: Debiasing Vision-Language Models with Adversarial Learning View paper
- [34] The dark side of dataset scaling: Evaluating racial classification in multimodal models View paper
- [35] See It from My Perspective: Diagnosing the Western Cultural Bias of Large Vision-Language Models in Image Understanding View paper
- [36] Worst of Both Worlds: Biases Compound in Pre-trained Vision-and-Language Models View paper
- [37] SB-Bench: Stereotype Bias Benchmark for Large Multimodal Models View paper
- [38] American== white in multimodal language-and-image ai View paper
- [39] User-vlm 360: Personalized vision language models with user-aware tuning for social human-robot interactions View paper
- [40] Measuring Social Biases in Grounded Vision and Language Embeddings View paper
- [41] Beyond the Surface: A Comprehensive Analysis of Implicit Bias in Vision-Language Models View paper
- [42] Vision Language Models are Biased View paper
- [43] Dear: Debiasing vision-language models with additive residuals View paper
- [44] Revealing and Reducing Gender Biases in Vision and Language Assistants (VLAs) View paper
- [45] DR.GAP: Mitigating Bias in Large Language Models using Gender-Aware Prompting with Demonstration and Reasoning View paper
- [46] Uncovering bias in large vision-language models at scale with counterfactuals View paper
- [47] Improving Bias Metrics in Vision-Language Models by Addressing Inherent Model Disabilities View paper
- [48] Bias in the Picture: Benchmarking VLMs with Social-Cue News Images and LLM-as-Judge Assessment View paper
- [49] Toward Socially Aware Vision-Language Models: Evaluating Cultural Competence Through Multimodal Story Generation View paper
- [50] Investigating Social Biases in Multimodal LLMs View paper
- [51] Words or Vision: Do Vision-Language Models Have Blind Faith in Text? View paper
- [52] Counterfactually Measuring and Eliminating Social Bias in Vision-Language Pre-training Models View paper
- [53] CLIP the Bias: How Useful is Balancing Data in Multimodal Learning? View paper
- [54] Red-Teaming for Inducing Societal Bias in Large Language Models View paper
- [55] Large language model as attributed training data generator: A tale of diversity and bias View paper
- [56] Can multimodal large language models enhance performance benefits among higher education students? An investigation based on the taskâ□□technology fit â□¦ View paper
- [57] Investors' acceptance and use of investment-based crowdfunding platforms: an integrated perspective View paper
- [58] Measuring Mechanistic Independence: Can Bias Be Removed Without Erasing Demographics? View paper
- [59] Assessing the Reliability of LLMs Annotations in the Context of Demographic Bias and Model Explanation View paper
- [60] Towards Alleviating the Object Bias in Prompt Tuning-based Factual Knowledge Extraction View paper
- [61] Selective Retrieval of Stimulus Information versus Thematic Judgments in Natural Language Inferences. View paper
- [62] Trustllm: Trustworthiness in large language models View paper
- [63] Multi-objective linguistic control of large language models View paper
- [64] Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned View paper
- [65] A Methodological Framework for Auditing Norm-Sensitive Behaviour in Large Language Models: Research Design for Employment Contexts View paper
- [66] Understanding Large Language Model Vulnerabilities to Social Bias Attacks View paper
- [67] Silenced Biases: The Dark Side LLMs Learned to Refuse View paper
- [68] Benchmarking Adversarial Robustness to Bias Elicitation in Large Language Models: Scalable Automated Assessment with LLM-as-a-Judge View paper

- [69] Bias Beyond Demographics: Probing Decision Boundaries in Black-Box LVLMs via Counterfactual VQA View paper
- [70] From Narrow Unlearning to Emergent Misalignment: Causes, Consequences, and Containment in LLMs View paper
- [71] One Model for All: Multi-Objective Controllable Language Models View paper