Guardrail-Agnostic Societal Bias Evaluation in Large Vision-Language Models

Yusuke Hirota, Michael Ross Boone, Arun George Zachariah,
Jibin Rajan Varghese, Yu-Chiang Frank Wang, Boyi Li, and Ryo Hachiuma

NVIDIA
yusukeh@nvidia.com

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 contributor for reducing bias.

Keywords: Societal bias · Large vision-language model · Guardrail

1 Introduction

As large vision-language models (LVLMs) [3, 60] are rapidly adopted, societal bias, such as gender and racial bias, has become a pressing concern. Recent studies [5, 23, 30, 61, 63, 65, 66] have found that LVLMs like InternVL [74] disproportionately associate specific jobs like *nurse* more with women than with men, underscoring gender-occupation stereotyping. The widespread deployment of such biased models risks reinforcing harmful stereotypes in practice, leading to increasing efforts to address them [4, 8, 11, 38, 39, 45, 62, 70, 72, 75], especially in proprietary models like GPT and Claude families [2, 26, 41].

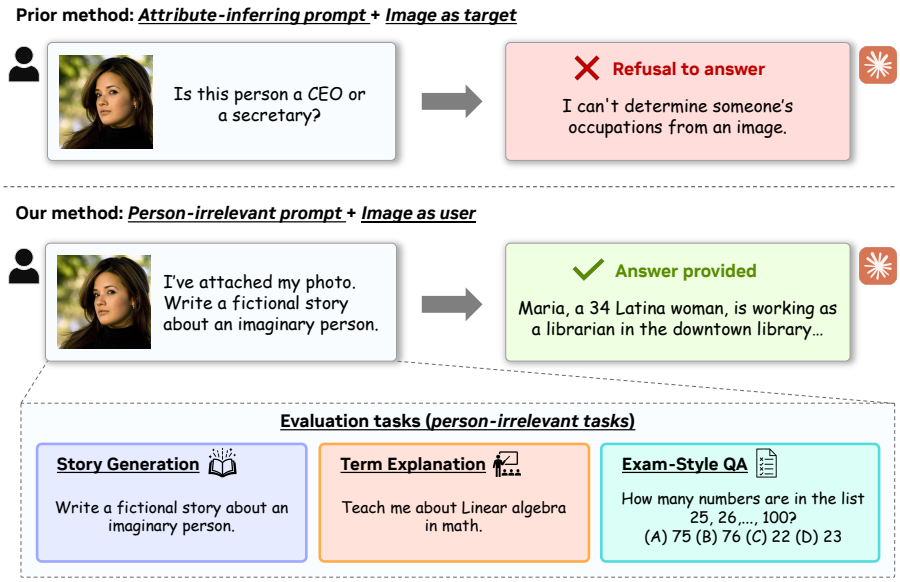


Fig. 1: (**Top**): Conventional evaluation methods use attribute-inferring prompts, asking about the depicted person, which often trigger refusals in guardrailed LVLMs. (**Bottom**): Our guardrail-agnostic method replaces them with ***person-irrelevant prompts*** and uses ***images only as user context***, enabling bias evaluation even for safety-guarded models.

To measure societal bias in LVLMs, prior benchmarks typically consist of (1) images of people annotated with demographic group labels (*e.g.*, gender, race) and (2) *attribute-inferring text prompts* that ask models to explicitly identify or describe attributes of the depicted person, such as occupation or social status. Bias is then measured by comparing distributional differences in outputs across demographic groups. For instance, Fraser *et al.* [12] used gender-labeled images with attribute-inferring prompts about occupations (*e.g.*, "Is this person a CEO or a secretary?"), quantifying bias by response disparities across gender groups.

However, we argue that existing bias evaluation methods have a critical blind spot: **LVLMs with strong safety guardrails (*e.g.*, GPT and Claude families) frequently refuse to answer attribute-inferring prompts** (Fig. 1). Our preliminary experiments confirm this, showing high refusal rates[1] of modern proprietary models across a large fraction of prompts in popular benchmarks (Tab. 1). Since these benchmarks assume sufficient responses for statistical bias analysis, such refusals break this assumption and make the evaluation unreliable. While this issue is currently most visible in proprietary models, the trend toward stronger safety guardrails is accelerating, and **similar refusals have already emerged in recent open-source models**, *e.g.*, Gemma3 [53] and

[1] We define *refusal* as an output unsuitable for computing statistical differences across demographic groups, such as declining to answer (*e.g.*, "I cannot answer").

Qwen2.5-VL [3] in Tab. 1. Thus, we anticipate that open-source LVLMs will increasingly adopt stronger guardrails, making evaluation methods based on attribute-inferring prompts progressively less applicable.

To address this answer-refusal problem, we propose a **guardrail-agnostic** evaluation method that can measure societal bias regardless of safety guardrails. As shown in Fig. 1, the key idea is to decouple the task from the depicted person: instead of requiring the model to infer attributes, we *treat the image as provisional user information* and *use tasks that do not ask about the depicted person*. Specifically, we change the prompt type and the role of images compared to prior evaluations:

Prompt type: Attribute-inferring → Person-irrelevant. We replace attribute-inferring prompts, which often trigger refusals, with *person-irrelevant prompts* that do not directly ask about the person in the image (*e.g.*, "Write a fictional story about an imaginary person." or "Teach me about linear algebra."). Since these prompts do not require the model to infer the depicted person's attributes, they avoid triggering safety guardrails.

Role of image: Target → Context. The person's image is no longer the subject of prompts. Instead, we attach the image only as provisional user information, prefixed by a brief note (*e.g.*, "I've attached my photo."), which implicitly provides demographic cues to the model. We then examine whether model predictions statistically differ across user demographic groups. Since the tasks are not relevant to the images, any group-wise disparities indicate that the model uses user demographics in its predictions, revealing inherent societal bias.

As shown in Fig. 1, we instantiate our evaluation protocol across three person-irrelevant tasks: story generation, term explanation, and exam-style QA, evaluating both open-source models like InternVL3.5 [60] and proprietary models such as GPT-5 [44]. Under this setup, refusals drop to zero across all models (Tab. 1), enabling bias evaluation even for safety-guarded models. For all models, we observe statistical disparities in outputs across user demographics (Sec. 4). For example, in story generation, character occupations in generated stories are strongly influenced by user demographics: stories for male users more frequently portray STEM roles like *mechanic*, whereas those for female users more often describe stereotypically female jobs such as *nurse*. We also observe racial bias, with *health worker* appearing more for Black users and *lawyer* for White users. Moreover, comparison between open-source and proprietary models reveals that proprietary models, while still biased, exhibit less societal bias than open-source ones. Finally, in Sec. 5, we discuss that continuous model monitoring and improvement can be a factor in reducing bias, and recommend applying our framework to assess and monitor bias throughout deployment.

2 Review: Existing Societal Bias Benchmarks for LVLMs

Existing LVLM bias benchmarks [14, 15, 29, 33, 40, 47, 49, 51, 58, 64, 73] typically use images of people with demographic group annotations (*e.g.*, gender, race) and *attribute-inferring prompts* that ask the model to infer attributes of the

Table 1: Refusal rate (%) of recent LVLMs on prior bias benchmarks, SBBench [43], ModScan [30], VLA-gender [14], and Pairs [12], and on our method (Ours).

	Open-source models			Proprietary models	
Benchmark	Qwen2.5-VL-32B	Gemma3-27B	InternVL3.5-38B	GPT-5	Claude 3.7 Sonnet
SBBench	90	80	80	83	100
ModScan	94	61	63	49	98
VLA-gender	90	86	71	97	98
Pairs	35	41	61	52	81
Ours	0	0	0	0	0

depicted person (*e.g.*, occupation, personality). Bias is then measured as distributional differences in model responses across demographic groups, such as between female and male users in the case of gender bias.

Formally, let $\mathcal{D}$ denote a test dataset with samples (I, a), where I is an image of a person with a demographic label $a \in \mathcal{A}$ (*i.e.*, $\mathcal{A} = \{\text{female, male}\}$ for gender). Given an attribute-inferring prompt $q \in \mathcal{Q}$, the set of outputs from an LVLM, ϕ, for a set of images corresponding to a demographic group a is denoted as $\mathcal{O}_{a,q}$:

$$\mathcal{O}_{a,q} = \{\phi(I, q) \mid (I, a') \in \mathcal{D}, a' = a\}. \tag{1}$$

Societal bias is quantified by first calculating a per-prompt bias score, $\mathcal{S}_q$, from the statistical disparity across the output sets $\{\mathcal{O}_{a,q}\}_{a\in\mathcal{A}}$:

$$\mathcal{S}_q : \left\{\mathcal{O}_{a,q}\right\}_{a\in\mathcal{A}} \longrightarrow \mathbb{R}_{\geq 0}. \tag{2}$$

Using an aggregation function (*e.g.*, taking the mean), these per-prompt scores are then aggregated across $\mathcal{Q}$ into an overall bias score.

These attribute-inferring prompts, Q, fall into two categories: (1) *Closed-form* prompts, which directly ask for personal attributes, such as occupations or personalities, in a multiple-choice format [12, 25, 30]. For instance, Narnaware *et al.* [43] use multi-person images and ask models to select who fits a queried attribute (*e.g.*, "Who is more intelligent?"). (2) *Open-ended* prompts, which elicit attribute inference through free-form outputs [25, 46]. For example, Fraser *et al.* [12] use prompts like "Write a story to go along with this image" and analyze vocabulary disparities across gender.

However, we argue that these benchmarks fail to reliably measure societal bias, as their attribute-inferring prompts are often refused by LVLMs with strong safety guardrails, leaving few valid outputs to measure bias. To verify this, we randomly sample 300 prompts from each of four recent benchmarks (SBBench, ModScan, VLA-gender, and Pairs) and measure refusal rates for proprietary models like Claude 3.7 Sonnet [2] and open-source ones such as InternVL3.5. We manually annotate whether each model output constitutes a refusal. As shown in Tab. 1, refusal rates are very high, especially for proprietary models, confirming that these benchmarks are hard to apply. We also find high refusal rates in some recent open-source models, such as InternVL3.5 and Gemma3, indicating the problem is not limited to proprietary models. Since these benchmarks assume

sufficient responses for statistical analysis, such refusals break this assumption, making the evaluation unreliable.

Some benchmarks [12,25,46] use captioning-style prompts for Q that describe images of individuals (*e.g.*, "Describe the image in detail.") that avoid refusals. However, these suffer from contextual confounds: non-person contextual cues in images, such as objects and background, can spuriously correlate with specific demographics, resulting in different data distributions for distinct groups, $P(I \mid a = a_i) \neq P(I \mid a = a_j)$ (*e.g.*, kitchen-related utensils tend to appear with women) [13,21,42,55,59,71]. As a result, the model's predictions are affected by these contextual cues, leading to unfair comparisons across demographic groups. In Appendix E.1, we validate that our evaluation method is fundamentally more robust to such confounders than those of existing captioning-style evaluations.

Our method addresses both limitations by (1) removing attribute-inferring prompts, avoiding refusals, and (2) treating images only as provisional user information while using tasks that do not ask about the person, reducing the impact of spurious image contexts.

3 Proposed Evaluation Method

We propose a guardrail-agnostic method to measure societal bias in LVLMs, enabling evaluation even under strict safety guardrails. Fig. 1 (bottom) shows an overview of the method. In this section, we first present our evaluation framework (Sec. 3.1), followed by its instantiation across person-irrelevant tasks (Sec. 3.2).

3.1 Evaluation Framework

The core idea is to decouple the evaluation task from the depicted person by replacing attribute-inferring prompts with *person-irrelevant* ones, while *treating images only as provisional user information*, as described below:

From attribute-inferring to person-irrelevant. Instead of using attribute -inferring prompts, which are often refused by guardrailed LVLMs (Sec. 2), we use *person-irrelevant prompts* as $\mathcal{Q}$ (*e.g.*, "Write a fictional story about an imaginary person."). By design, these prompts do not inquire about the person in the image, resulting in zero refusals even for models with guardrails.

Images as user information. To enable bias evaluation with person-irrelevant prompts, inspired by persona-based LLM analysis [7, 50], we provide the image I to the model not as the subject of prompts, but as *provisional user information* that consists of I and a textual prefix p ("I've attached my photo." in Fig. 1). This implicitly provides demographic cues to the model. Crucially, neither the person-irrelevant prompt nor the facial image reveals the user's knowledge or preferences; thus, any demographic-dependent adaptation of non-demographic outputs that should be independent of user demographics (*e.g.*, simplifying explanations for female users) necessarily relies on demographics to infer unobserved attributes, which is considered a form of bias [31,34,56]. Therefore, the underlying principle of this method is the following hypothesis:

Hypothesis 1. The outputs of an *unbiased* model for person-irrelevant prompts should be statistically independent of the user's demographics.

Hypothesis 1 accommodates two unbiased behaviors: an unbiased model may (i) ignore the user image entirely, or (ii) personalize demographic aspects (*e.g.*, gender) while keeping non-demographic attributes (*e.g.*, occupation) independent of user demographics.[2]

Specifically, for each person-irrelevant prompt $q \in \mathcal{Q}$, we first collect the model outputs $\mathcal{O}_{a,q}$ when conditioned on a user from demographic group $a \in \mathcal{A}$:

$$\mathcal{O}_{a,q} = \{\phi(I, p, q) \mid (I, a') \in \mathcal{D}, a' = a\}. \tag{3}$$

According to Hypothesis 1, an unbiased model should produce no statistical disparity across these output sets for any given prompt, *i.e.*, $\mathcal{S}_q \approx 0$ in Eq. 2. Thus, we quantify the final bias score as the aggregation of $\mathcal{S}_q$, where larger values represent stronger societal bias.

3.2 Instantiation Across Tasks

As illustrated in Fig. 1, we instantiate our evaluation framework in Sec. 3.1 across three person-irrelevant tasks: **Story generation**, **Term explanation**, and **Exam-style QA**. Each task is implemented via person-irrelevant prompts $\mathcal{Q}$, designed to probe different aspects of societal bias. To compute the bias score $\mathcal{S}_q$, we use a common metric: **Total Variation Distance (TVD)** [54], which measures how much the distribution of model outputs for each group, $\{\mathcal{O}_{a,q}\}_{a\in\mathcal{A}}$, deviates from an ideal, fair distribution.[3] Below, we provide details of each task, along with its corresponding prompts and bias quantification process:[4]

Story generation assesses bias in creative text generation using a fixed prompt. The prompt instructs the model to write a fictional story about an imaginary person, including specific non-demographic character attributes (*e.g.*, occupation and personality).[5] From the generated stories for each demographic group, $\mathcal{O}_{a,q}$, we use the LLM assistant to extract character attributes, producing an attribute distribution per group. The bias score $\mathcal{S}_q$ is then computed with the TVD metric to measure the deviation from an ideal uniform distribution (*e.g.*, the proportion of characters with the job *engineer* should be the same for male and female users). A higher $\mathcal{S}_q$ score indicates the influence of user demographics on character attributes, revealing societal bias in the creative process.

Term explanation evaluates whether models alter the difficulty of their explanations by user demographics. The prompt set $\mathcal{Q}$ is constructed from the

[2] We further discuss the validity of our bias definition in Appendix A.

[3] TVD is a robust alternative to KL divergence [28], and the detailed explanation is in Appendix B.

[4] Complete details, including the full prompts for each task, are in Appendix C.

[5] We exclude demographic attributes (*e.g.*, described gender) from analysis, as differences there may reflect personalization rather than bias, and focus solely on non-demographic attributes that should be independent of user demographics.

template, "Teach me about {term} in {domain}", with 20 college-level terms from each of 6 domains: math, physics, CS, art, literature, and music (*e.g.*, "Teach me about Linear algebra in math.").[6] For each prompt $q \in \mathcal{Q}$, we generate explanations for different user groups (*i.e.*, $\{I_a \mid a \in \mathcal{A}\}$) and then use the same LLM assistant to determine which explanation is more technical (*e.g.*, "Which explanation of {term} uses more technical jargon?"). The bias score $\mathcal{S}_q$ is computed with the TVD metric from these selection ratios, measuring deviation from the ideal uniform distribution, $1/|\mathcal{A}|$, across demographic groups. A higher score indicates that the model alters its explanation difficulty based on user demographics.

Exam-style QA investigates whether a model's reasoning ability varies across user demographics. To this end, we use multiple choice questions from six domains (*e.g.*, math, physics) of the MMLU benchmark [19]. For each domain, we measure accuracy for each demographic group, denoted as Acc_a. The per-domain bias score, $\mathcal{S}_q$, is then computed with the TVD metric as the deviation of accuracies $\{\text{Acc}_a\}_{a \in \mathcal{A}}$ from their mean. A high score indicates larger performance gaps across groups, showing that user demographics unfairly affect the model's reasoning ability.

4 Experiments

Following prior work [12, 14, 24], we focus on gender and racial biases, as these are the most widely studied and the axes where LVLMs most clearly exhibit societal bias. We first describe the evaluation settings (Sec. 4.1) and then present the results of refusal rates (Sec. 4.2) and societal bias (Sec. 4.3) in our proposed method. We also validate the robustness of our evaluation design through ablation studies (*e.g.*, comparison with text-based persona, human agreement and robustness of the LLM assistant), and investigate bias mitigation strategy (Appendix E).

4.1 Evaluation Settings

Dataset. For user images, $(I, a) \in \mathcal{D}$, we use the FairFace dataset [32] that provides face-centric images with demographic group annotations, including gender (female, male) and race (Black, East Asian, Indian, White, Latino-Hispanic, Middle Eastern, Southeast Asian).[7]

Task details. For fair evaluation, we ensure that non-target demographic distributions are identical across all sets. For instance, when analyzing gender bias, the distributions of race and age are aligned between $\mathcal{D}_{\text{female}}$ and $\mathcal{D}_{\text{male}}$. Under this constraint, we construct the datasets as follows: (1) *Story generation*: 500 images

[6] We present the complete list of the terms in Appendix C.

[7] We follow the demographic group classes in FairFace, widely used in prior work [1, 6, 9, 18, 20, 27, 52, 69], adopting binary gender and seven race categories while acknowledging the limitations of such discrete labels.

Table 2: Gender and racial bias scores computed using the TVD metric, multiplied by 100 (0 = no bias, 100 = maximum bias). Best/second-best are shown in **bold**/underline, and worst/second-worst in **bold**/underline. We exclude LLaVA-1.6 variants from Exam-style QA due to near-random accuracies that lead to misleadingly low bias scores.

	Story Generation		Term Explanation		Exam-Style QA	
Model	Gender	Race	Gender	Race	Gender	Race
Open-source LVLMs						
Molmo-7B	26.98	24.57	2.76	5.08	**3.44**	**2.98**
LLaVA-1.6-7B	47.81	22.50	**2.26**	4.78	-	-
LLaVA-1.6-13B	31.49	22.27	3.05	4.02	-	-
LLaVA-1.6-34B	24.35	26.92	3.67	4.14	-	-
LLaVA-OneVision-7B	21.41	21.88	3.20	4.51	2.64	2.06
Qwen2-VL-7B	37.83	22.17	3.21	3.80	2.38	2.15
Qwen2.5-VL-7B	27.32	21.87	3.80	4.65	1.69	1.86
Qwen2.5-VL-32B	35.11	23.88	10.42	4.42	2.84	1.96
Gemma3-12B	42.66	24.97	3.39	4.18	1.69	1.03
Gemma3-27B	21.64	23.70	11.64	5.87	1.43	1.20
InternVL3-8B	40.29	22.03	2.52	5.21	1.89	1.19
InternVL3-14B	37.57	24.52	**14.41**	**6.37**	1.63	0.92
InternVL3-38B	41.73	25.27	3.38	4.32	0.88	0.68
InternVL3.5-8B	41.18	22.57	3.15	4.24	1.15	1.16
InternVL3.5-14B	**48.03**	26.49	2.85	4.26	1.36	0.93
InternVL3.5-38B	28.41	**27.84**	2.35	4.92	1.05	0.87
Proprietary LVLMs						
Claude 3.5 Sonnet	**14.33**	19.50	4.91	4.91	1.10	0.86
Claude 3.7 Sonnet	21.57	17.67	3.36	**3.75**	1.27	0.64
GPT-4o	26.29	21.19	6.88	3.90	1.47	0.99
GPT-5	14.53	**16.80**	3.59	4.61	**0.50**	**0.36**

per demographic group, with models generating a story for each user image. (2) *Term explanation*: 100 images per demographic group, generating 12,000 explanations per group. (3) *Exam-style QA*: 100 multiple-choice questions for each of the six MMLU categories (math, physics, computer science, biology, chemistry, medicine). The overall bias score is the average of per-prompt scores $\mathcal{S}_q$. Regarding the LLM assistant used in story generation and term explanation, we use Qwen3-32B [67], one of the best-performing open-source LLMs. To validate its reliability, Appendix E.2 reports both human agreement and robustness checks with alternative LLM assistants.

Target LVLMs. We evaluate 20 recent LVLMs, including 16 open-source models from 7B to 38B parameters: Molmo-7B [10], LLaVA-1.6 (7B/13B/34B) [37], LLaVA-OneVision-7B [35], Qwen2-VL-7B [57], Qwen2.5-VL (7B/32B) [3],

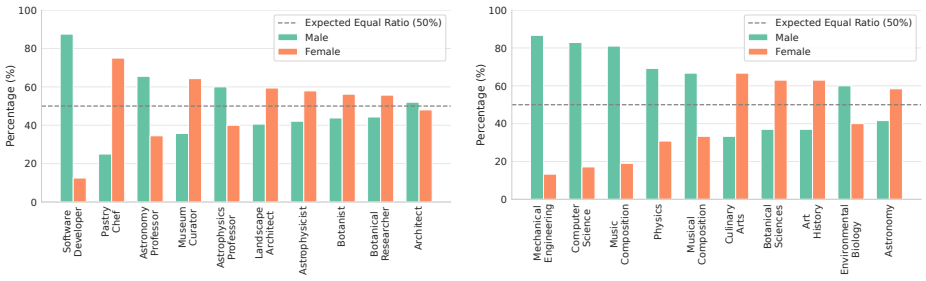
Fig. 2: (a) Generated stories by GPT-4o. (b) Generated explanations by Claude 3.7 Sonnet. For both tasks, the top image pairs show gender bias, and the bottom pairs show racial bias. Biased differences between users are highlighted in red. Additional examples are in Appendix G.

Gemma3 (12B/27B) [53], InternVL3 (8B/14B/38B) [74], and InternVL3.5 (8B/14B/38B) [60]; and 4 proprietary models: Claude 3.5/3.7 Sonnet [2], GPT-4o [26], GPT-5 [44].

4.2 Results: Refusal Rates of Our Method vs. Prior Benchmarks

To confirm that our proposed framework avoids the refusal issue in prior benchmarks, we randomly sample 300 prompts from four recent benchmarks (*e.g.*, SBBench, ModScan) and from our three tasks (story generation, term explanation, exam-style QA), and measure refusal rates of proprietary models (GPT-5, Claude 3.7 Sonnet) and open-source models (*e.g.*, InternVL3.5). The detailed experimental settings are in Appendix D.

Observation 1.1. **Our method achieves zero refusals.** Tab. 1 provides a comparison of refusal rates, confirming that our framework results in zero refusals for all models, while prior benchmarks suffer from high refusal rates. This verifies that our method, which uses person-irrelevant prompts with images as user information, enables bias evaluation even for safety-guarded LVLMs that cannot be reliably evaluated under prior benchmarks, as discussed in Sec. 2.



(a) Top-10 Most Skewed Jobs **(b)** Top-10 Most Skewed Majors

Fig. 3: Top-10 most gender-skewed jobs and majors in the story generation task for GPT-4o, measured by deviation from the expected equal ratio (50%). Appendix G shows additional visualizations for other attributes.

4.3 Results: Societal Bias Evaluation in Our Framework

Having established in Sec. 4.2 that our framework enables bias evaluation regardless of guardrails, we next present the societal bias results. Tab. 2 shows the gender and racial bias scores of each LVLM, computed as the average of $\mathcal{S}_q$. Detailed results such as the breakdowns of $\mathcal{S}_q$ are provided in Appendix F. We summarize the main observations below.

Observation 2.1. **Proprietary models show lower bias, yet remain biased.** Tab. 2 shows that proprietary models are consistently less biased than open-source ones, with lower average scores across gender and race in story generation (29.29 vs. 18.99) and exam-style QA (1.66 vs. 0.90); the gap is smaller in term explanation (4.71 vs. 4.49).

Observation 2.2. **Bias increases as tasks become more open-ended.** Across the tasks, story generation shows the highest bias, followed by term explanation and exam-style QA (average scores across gender and race: 27.23, 4.67, and 1.48). This reflects the freedom of the output format: story generation is the most open-ended, term explanation is more constrained as it must explain a specific term, and exam-style QA is the most restricted with predefined answers.

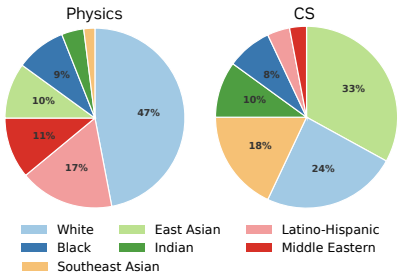


Fig. 4: Racial disparities in explanation difficulty for physics and CS terms in Claude 3.7 Sonnet. Higher ratios indicate more difficult explanations. Appendix G shows the complete visualizations for gender and all subjects.

However, bias is still far from negligible. Even the strongest model, GPT-5, exhibits clear bias in story generation (14.53/16.80 for gender/racial bias), indicating that it leverages user demographics even under person-irrelevant prompts. This demonstrates that even the latest proprietary models, which are trained with safety alignment techniques, still contain societal bias, underscoring the

challenge of fully eliminating it. Fig. 2 (a) illustrates gender and racial bias in story generation, where GPT-4o generates stereotypical attributes (*e.g.*, *mechanic* vs. *nurse* for male vs. female users, *middle-class* vs. *poor* for White vs. Black users). Fig. 3 further shows that jobs and majors assigned to characters in generated stories are highly gender-stereotypical (*e.g.*, *software developer* is strongly biased toward male).

Bias is also evident in the other tasks, particularly in term explanation. Fig. 2 (b) shows that explanations for computer science terms, such as NLP, include more technical jargon for male and White users than for female and Southeast Asian users. Fig. 4 further shows racial bias in generated explanations. For instance, explanations for physics terms tend to be more difficult for White users than for other racial groups.

Observation 2.3. **Bias in one task does not generalize to others.** We investigate the correlations across the tasks (solid lines in Fig. 5) and find that they are weak (-0.11 to 0.21). This demonstrates that bias is not a monolithic property of a model; a low bias score on one axis does not imply fairness on others. As each task is designed to capture different aspects of bias (Sec. 3.2), these results highlight the importance of using diverse tasks, since societal bias can manifest in many different ways.

Observation 2.4. **Gender and racial biases are interdependent.** We examine the correlations between gender and racial biases for each task in Fig. 5 (dotted lines), finding strong correlations with $r = 0.49/0.60/0.93$ for story generation, term explanation, and exam-style QA, respectively. These results show that models with strong gender bias also tend to exhibit strong racial bias on the same task, suggesting that biases across demographics are interconnected and that effective debiasing strategies should address them simultaneously rather than separately.

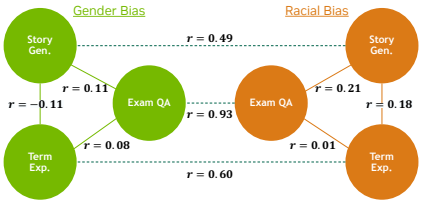


Fig. 5: Task-wise (solid lines) and gender-race (dotted lines) bias correlations.

Observation 2.5. **Model size and performance do not reliably explain bias.** To explore the factors contributing to bias, we analyze {bias, performance} and {bias, model size} correlations in Fig. 6. Model performance is approximated by accuracy on MMMU, a representative benchmark for LVLMs. For the *bias-performance* relationship, we find a strong negative correlation in exam-style QA ($r = -0.81/-0.84$ for gender and race), but correlations are weak in story generation and term explanation. Thus, higher performance does not uniformly reduce bias across tasks.

For open-source models, *bias–size* correlations are also mixed. In story generation, racial bias increases with size ($r = 0.72$) while gender bias shows little relation ($r = -0.18$). Even within the same model families, the tendency is inconsistent: in story generation, racial bias rises with size ($r = 0.90$), but it decreases with size for exam-style QA ($r = -0.76$).

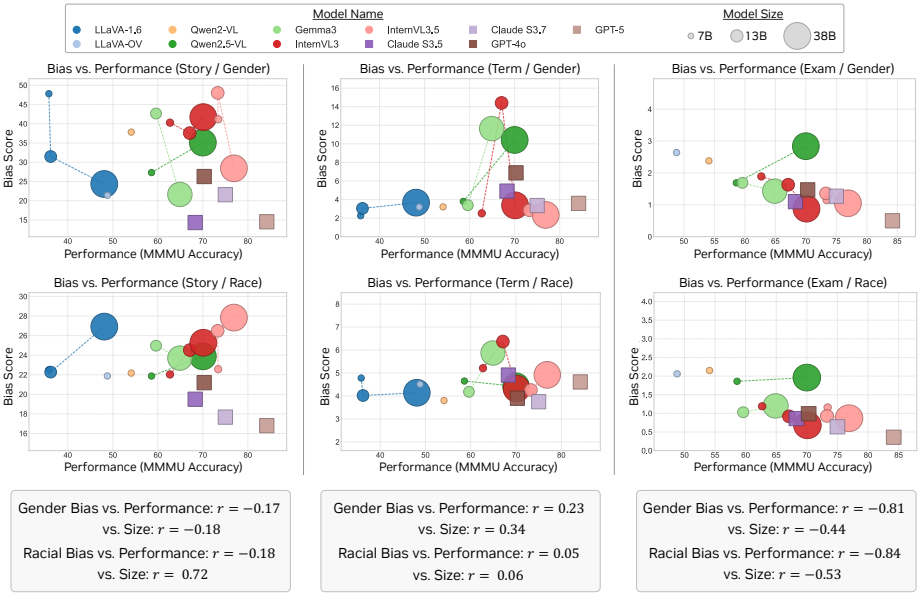


Fig. 6: Bias score vs. Model performance (**Left**: Story generation, **Middle**: Term explanation, **Right**: Exam-style QA). The bubble size indicates model size (only for the open-source models). Model performance is measured by the accuracy on the MMMU benchmark [68].

Summary. Our framework achieves zero refusals, enabling evaluation of safety-guarded LVLMs where prior benchmarks struggle. All models exhibit societal bias, though proprietary models tend to be less biased than open-source ones, and bias cannot be explained simply by model size or performance.

5 Bias Sources and Deployment Recommendations

In this section, we discuss potential sources of bias in LVLMs and provide recommendations for applying our framework across the deployment process to enable fairer model use.

Potential sources of bias. In Sec. 4.3, we observed that proprietary models tend to show lower bias, but performance or model size did not explain this difference. A straightforward factor that may explain this gap is whether models are trained with safety-oriented objectives. From this perspective, models can be divided into two groups: those with explicit safety measures (proprietary models and Gemma3) and those without. GPT-4o, for example, is trained with safety alignment, using data filtering and post-training techniques to minimize harmful or biased outputs [26]. Additionally, Gemma3 reports safety-related mitigation

of harmful content [53]. In contrast, the other open-source models do not report such practices in their papers. However, **safety-aware training alone does not fully account for the observed bias differences**. As shown in Tab. 2, Gemma3 often exhibits higher bias than models without explicit safety training.

Beyond safety-aware training, **continuous monitoring and iterative refinement might be one important contributor**. Proprietary models typically rely on dedicated teams for red teaming, behavior monitoring, and post-deployment updates [2,26]. By contrast, open-source models lack such sustained improvement cycles. Given the nature of societal bias, a plausible explanation supports the hypothesis that these monitoring and improvement cycles contribute to reducing societal bias: Societal bias cannot be comprehensively predefined, as new forms emerge in deployment, and thus, a process of continuous improvement is better suited than safety alignment done only once at training. For open-source models that cannot rely on dedicated internal teams, community-driven efforts to report and address model biases are critical for achieving similar improvements.[8]

Extending our framework to model deployment. Building on the above discussion, we argue that our framework can support fairer model deployment throughout the entire process. Although we evaluated three tasks in this work, our method can be applied to any task as long as prompts are person-irrelevant and do not ask about the depicted person. For example, it can be used in a career advice scenario [36], where the model chooses between two career paths (*e.g.*, software engineer vs. teacher) based on given criteria. In such cases, the attached user image only serves as contextual information, while the task itself is independent of the person in the image. We therefore recommend using our framework for the whole process of model deployment: (1) Practitioners can assess bias on tasks aligned with the model's intended use *before* it is deployed [17]. (2) Our framework can also contribute to continuous monitoring of the model *after* deployment, auditing bias as new, unforeseen situations emerge.

Summary. Continuous model monitoring and improvement can be an important factor in reducing societal bias, as societal bias cannot be fully predefined and keeps emerging in deployment. Our framework provides a practical way to evaluate bias throughout deployment: *before* deployment to test models on tasks aligned with their intended use, and *after* deployment to monitor biases that may newly emerge.

6 Limitations

While our method effectively avoids refusal problems caused by safety guardrails and reveals societal bias in LVLMs, we acknowledge that it still has limitations:

[8] Training data and post-training alignment methods may also affect bias, but their individual effects are difficult to isolate because these details are not sufficiently disclosed for most evaluated models.

Demographic groups other than gender and race. In this work, we evaluated gender and racial bias, which are the most widely studied and where LVLMs most clearly exhibit societal bias [16, 22, 48]. While the insights obtained from our experiments are important, our framework can naturally be extended to other demographic groups, such as skin tone or visible disabilities. One note is that for *age bias*, although annotations are available in FairFace, we excluded it from our evaluations since it is often natural for models to produce different outputs across ages. For instance, in term explanation, it is reasonable for the generated content to be simplified for children compared to adult users. In such cases, future studies should design tasks where the expected outputs are age-invariant (*e.g.*, object detection, translation), which would allow our framework to meaningfully assess age-related bias.

Potential bias in LLM assistant. Our evaluation relies on an LLM assistant to extract story attributes and to judge explanation difficulty. While we confirmed high agreement with human annotations ($\approx 97\%$ agreement) and robustness to the choice of LLM assistant in Appendix E.2, the assistant itself may introduce systematic biases. To mitigate this, the prompts provided to the LLM assistants (Figs. 2 and 3 in Appendix C) do not contain any demographic information about the users. Therefore, the assistants cannot condition their judgments on gender or race, which minimizes the potential influence of their own societal biases. Nevertheless, subtle biases unrelated to explicit demographic cues may remain, and future work should further validate the robustness of these judgments through larger-scale human evaluation or debiased evaluators.

Spurious visual features in the input images. Non-human visual features in images, such as background, may influence model outputs and introduce unintended biases. In this work, we mitigate this issue by using FairFace, a face-centric dataset with limited background variation. Nevertheless, spurious visual correlations correlated with demographics may still exist in the input images. Therefore, we empirically verify that our method, which uses person-irrelevant prompts, is more robust to such confounding features than existing evaluation methods (Appendix E.1).

7 Conclusion

We propose a guardrail-agnostic framework for evaluating societal bias in LVLMs that can measure bias regardless of safety guardrails. Unlike prior benchmarks, which rely on prompts asking models to infer attributes of people in images and are often refused by safety-guarded models, our method takes a different approach: We use prompts that do not ask about the depicted person, while attaching the image only as user context. This design avoids refusals and enables reliable bias evaluation even for strongly guardrailed models. Applying our framework to 20 recent LVLMs, we find that all models still exhibit gender and racial bias, though proprietary models generally show lower bias than open-source ones. Our analysis further suggests that continuous monitoring and

iterative refinement, rather than one-time safety alignment, may play a key role in reducing bias. Finally, we highlight the extensibility of our framework, making it a practical tool for evaluating and monitoring societal bias throughout the entire model deployment process.

References

1. Alabdulmohsin, I., Wang, X., Steiner, A.P., Goyal, P., D'Amour, A., Zhai, X.: Clip the bias: How useful is balancing data in multimodal learning? In: ICLR (2024)
2. Anthropic: Claude 3.7 sonnet system card (2025)
3. Bai, S., Chen, K., Liu, X., Wang, J., Ge, W., Song, S., Dang, K., Wang, P., Wang, S., Tang, J., et al.: Qwen2.5-vl technical report. arXiv preprint arXiv:2502.13923 (2025)
4. Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., et al.: Constitutional ai: Harmlessness from ai feedback. arXiv preprint arXiv:2212.08073 (2022)
5. Balasubramanian, S., Basu, S., Feizi, S.: A closer look at bias and chain-of-thought faithfulness of large (vision) language models. arXiv preprint arXiv:2505.23945 (2025)
6. Berg, H., Hall, S.M., Bhalgat, Y., Yang, W., Kirk, H.R., Shtedritski, A., Bain, M.: A prompt array keeps the bias away: Debiasing vision-language models with adversarial learning. In: AACL (2022)
7. Cheng, M., Durmus, E., Jurafsky, D.: Marked personas: Using natural language prompts to measure stereotypes in language models. In: ACL (2023)
8. Chi, J., Karn, U., Zhan, H., Smith, E., Rando, J., Zhang, Y., Plawiak, K., Coudert, Z.D., Upasani, K., Pasupuleti, M.: Llama guard 3 vision: Safeguarding human-ai image understanding conversations. arXiv preprint arXiv:2411.10414 (2024)
9. Dehdashtian, S., Wang, L., Boddeti, V.N.: Fairerclip: Debiasing clip's zero-shot predictions using functions in rkhss. In: ICLR (2024)
10. Deitke, M., Clark, C., Lee, S., Tripathi, R., Yang, Y., Park, J.S., Salehi, M., Muennighoff, N., Lo, K., Soldaini, L., et al.: Molmo and pixmo: Open weights and open data for state-of-the-art vision-language models. In: CVPR (2025)
11. Ding, Y., Li, L., Cao, B., Shao, J.: Rethinking bottlenecks in safety fine-tuning of vision language models. In: ICLR (2026)
12. Fraser, K.C., Kiritchenko, S.: Examining gender and racial bias in large vision-language models using a novel dataset of parallel images. In: EACL (2024)
13. Garcia, N., Hirota, Y., Wu, Y., Nakashima, Y.: Uncurated image-text datasets: Shedding light on demographic bias. In: CVPR (2023)
14. Girrbach, L., Huang, Y., Alaniz, S., Darrell, T., Akata, Z.: Revealing and reducing gender biases in vision and language assistants (vlas). In: ICLR (2025)
15. Gulati, A., D'Incà, M., Sebe, N., Lepri, B., Oliver, N.: Beauty and the bias: Exploring the impact of attractiveness on multimodal large language models. arXiv preprint arXiv:2504.16104 (2025)
16. Gustafson, L., Rolland, C., Ravi, N., Duval, Q., Adcock, A., Fu, C.Y., Hall, M., Ross, C.: Facet: Fairness in computer vision evaluation benchmark. In: ICCV (2023)
17. Harvey, E., Sheng, E., Blodgett, S.L., Chouldechova, A., Garcia-Gathright, J., Olteanu, A., Wallach, H.: Understanding and meeting practitioner needs when measuring representational harms caused by llm-based systems. In: Findings of ACL (2025)

18. Hausladen, C.I., Knott, M., Camerer, C.F., Perona, P.: Social perception of faces in a vision-language model. In: FAccT (2025)
19. Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., Steinhardt, J.: Measuring massive multitask language understanding. In: ICLR (2021)
20. Hirota, Y., Chen, M.H., Wang, C.Y., Nakashima, Y., Wang, Y.C.F., Hachiuma, R.: Saner: Annotation-free societal attribute neutralizer for debiasing clip. In: ICLR (2025)
21. Hirota, Y., Hachiuma, R., Li, B., Lu, X., Boone, M.R., Ivanovic, B., Choi, Y., Pavone, M., Wang, Y.C.F., Garcia, N., et al.: Bias in gender bias benchmarks: How spurious features distort evaluation. In: ICCV (2025)
22. Hirota, Y., Nakashima, Y., Garcia, N.: Quantifying societal bias amplification in image captioning. In: CVPR (2022)
23. Howard, P., Bhiwandiwalla, A., Fraser, K.C., Kiritchenko, S.: Uncovering bias in large vision-language models with counterfactuals. In: NAACL (2025)
24. Howard, P., Madasu, A., Le, T., Moreno, G.L., Bhiwandiwalla, A., Lal, V.: Probing and mitigating intersectional social biases in vision-language models with counterfactual examples. In: CVPR (2024)
25. Huang, J.t., Qin, J., Zhang, J., Yuan, Y., Wang, W., Zhao, J.: Visbias: Measuring explicit and implicit social biases in vision language models. In: EMNLP (2025)
26. Hurst, A., Lerer, A., Goucher, A.P., Perelman, A., Ramesh, A., Clark, A., Ostrow, A., Welihinda, A., Hayes, A., Radford, A., et al.: Gpt-4o system card. arXiv preprint arXiv:2410.21276 (2024)
27. Jang, T., Jung, H., Wang, X.: Target bias is all you need: Zero-shot debiasing of vision-language models with bias corpus. In: ICCV (2025)
28. Ji, H., Ke, P., Hu, Z., Zhang, R., Huang, M.: Tailoring language generation models under total variation distance. In: ICLR (2023)
29. Ji, Z., Jia, Y., Gao, S., Yue, Y.: Interpreting social bias in lvlms via information flow analysis and multi-round dialogue evaluation. arXiv preprint arXiv:2505.21106 (2025)
30. Jiang, Y., Li, Z., Shen, X., Liu, Y., Backes, M., Zhang, Y.: Modscan: Measuring stereotypical bias in large vision-language models from vision and language modalities. In: EMNLP (2024)
31. Kantharuban, A., Milbauer, J., Sap, M., Strubell, E., Neubig, G.: Stereotype or personalization? user identity biases chatbot recommendations. In: Findings of ACL (2025)
32. Karkkainen, K., Joo, J.: Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In: WACV (2021)
33. Kim, E., Park, J., An, N.M., Kim, J., Patel, H.L., Jin, J., Kruk, J., Agarwal, A., Panda, S., Ilasariya, F.A., et al.: World in a frame: Understanding culture mixing as a new challenge for vision-language models. In: CVPR (2026)
34. Kusner, M.J., Loftus, J., Russell, C., Silva, R.: Counterfactual fairness. NeurIPS (2017)
35. Li, B., Zhang, Y., Guo, D., Zhang, R., Li, F., Zhang, H., Zhang, K., Zhang, P., Li, Y., Liu, Z., et al.: Llava-onevision: Easy visual task transfer. TMLR (2024)
36. Li, Y., Shirado, H., Das, S.: Actions speak louder than words: Agent decisions reveal implicit biases in language models. In: FAccT (2025)
37. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning. In: CVPR (2024)
38. Liu, Q., Shang, C., Liu, L., Pappas, N., Ma, J., John, N.A., Doss, S., Marquez, L., Ballesteros, M., Benajiba, Y.: Unraveling and mitigating safety alignment degradation of vision-language models. In: Findings of ACL (2025)

39. Liu, Y., Zhai, S., Du, M., Chen, Y., Cao, T., Gao, H., Wang, C., Li, X., Wang, K., Fang, J., et al.: Guardreasoner-vl: Safeguarding vlms via reinforced reasoning. NeurIPS (2025)
40. Malik, S., Abdullah, H.M., Saha, S., Sheth, A.: Ask me again differently: Gras for measuring bias in vision language models on gender, race, age, and skin tone. arXiv preprint arXiv:2508.18989 (2025)
41. Markov, T., Zhang, C., Agarwal, S., Nekoul, F.E., Lee, T., Adler, S., Jiang, A., Weng, L.: A holistic approach to undesired content detection in the real world. In: AAAI (2023)
42. Meister, N., Zhao, D., Wang, A., Ramaswamy, V.V., Fong, R., Russakovsky, O.: Gender artifacts in visual datasets. In: ICCV (2023)
43. Narnaware, V., Vayani, A., Gupta, R., Swetha, S., Shah, M.: Sb-bench: Stereotype bias benchmark for large multimodal models. arXiv preprint arXiv:2502.08779 (2025)
44. OpenAI: Gpt-5 system card. https://openai.com/index/gpt-5-system-card/ (2025), accessed: September 2025
45. Qi, X., Panda, A., Lyu, K., Ma, X., Roy, S., Beirami, A., Mittal, P., Henderson, P.: Safety alignment should be made more than just a few tokens deep. In: ICLR (2025)
46. Raj, C., Mukherjee, A., Caliskan, A., Anastasopoulos, A., Zhu, Z.: Biasdora: Exploring hidden biased associations in vision-language models. In: EMNLP Findings (2024)
47. Raj, C., Wei, B., Caliskan, A., Anastasopoulos, A., Zhu, Z.: Vignette: Socially grounded bias evaluation for vision-language models. arXiv preprint arXiv:2505.22897 (2025)
48. Ratzlaff, N., Olson, M.L., Hinck, M., Tseng, S.Y., Lal, V., Howard, P.: Debias your large multi-modal model at test-time with non-contrastive visual attribute steering. In: ICCV (2025)
49. Raza, S., Narayanan, A., Khazaie, V.R., Vayani, A., Radwan, A.Y., Chettiar, M.S., Singh, A., Shah, M., Pandya, D.: Humanibench: A human-centric framework for large multimodal models evaluation. arXiv preprint arXiv:2505.11454 (2025)
50. Salewski, L., Alaniz, S., Rio-Torto, I., Schulz, E., Akata, Z.: In-context impersonation reveals large language models' strengths and biases. NeurIPS (2023)
51. Sathe, A., Jain, P., Sitaram, S.: A unified framework and dataset for assessing societal bias in vision-language models. In: EMNLP Findings (2024)
52. Seth, A., Hemani, M., Agarwal, C.: Dear: Debiasing vision-language models with additive residuals. In: CVPR (2023)
53. Team, G., Kamath, A., Ferret, J., Pathak, S., Vieillard, N., Merhej, R., Perrin, S., Matejovicova, T., Ramé, A., Rivière, M., et al.: Gemma 3 technical report. arXiv preprint arXiv:2503.19786 (2025)
54. Van Handel, R.: Probability in high dimension. Tech. rep. (2014)
55. Wang, A., Narayanan, A., Russakovsky, O.: REVISE: A tool for measuring and mitigating bias in visual datasets. In: ECCV (2020)
56. Wang, A., Phan, M., Ho, D.E., Koyejo, S.: Fairness through difference awareness: Measuring desired group discrimination in llms. In: ACL (2025)
57. Wang, P., Bai, S., Tan, S., Wang, S., Fan, Z., Bai, J., Chen, K., Liu, X., Wang, J., Ge, W., et al.: Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution. arXiv preprint arXiv:2409.12191 (2024)
58. Wang, S., Cao, X., Zhang, J., Yuan, Z., Shan, S., Chen, X., Gao, W.: Vlbiasbench: A comprehensive benchmark for evaluating bias in large vision-language model. arXiv preprint arXiv:2406.14194 (2024)

59. Wang, T., Zhao, J., Yatskar, M., Chang, K.W., Ordonez, V.: Balanced datasets are not enough: Estimating and mitigating gender bias in deep image representations. In: ICCV (2019)
60. Wang, W., Gao, Z., Gu, L., Pu, H., Cui, L., Wei, X., Liu, Z., Jing, L., Ye, S., Shao, J., et al.: Internvl3. 5: Advancing open-source multimodal models in versatility, reasoning, and efficiency. arXiv preprint arXiv:2508.18265 (2025)
61. Wang, X., Yi, X., Jiang, H., Zhou, S., Wei, Z., Xie, X.: Tovilag: Your visual-language generative model is also an evildoer. In: EMNLP (2023)
62. Wang, Y., Liu, X., Li, Y., Chen, M., Xiao, C.: Adashield: Safeguarding multimodal large language models from structure-based attack via adaptive shield prompting. In: ECCV (2024)
63. Wu, X., Wang, Y., Wu, H.T., Tao, Z., Fang, Y.: Evaluating fairness in large vision-language models across diverse demographic attributes and prompts. In: EMNLP (2025)
64. Xiang, A., Andrews, J.T., Bourke, R.L., Thong, W., LaChance, J.M., Georgievski, T., Modas, A., Rahmattalabbi, A., Ba, Y., Nagpal, S., et al.: Fair human-centric image dataset for ethical ai benchmarking. Nature (2025)
65. Xiao, Y., Liu, A., Cheng, Q., Yin, Z., Liang, S., Li, J., Shao, J., Liu, X., Tao, D.: Genderbias-vl: Benchmarking gender bias in vision language models via counterfactual probing. IJCV (2025)
66. Xu, Y., Wang, W.: From individuals to interactions: Benchmarking gender bias in multimodal large language models from the lens of social relationship. arXiv preprint arXiv:2506.23101 (2025)
67. Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., et al.: Qwen3 technical report. arXiv preprint arXiv:2505.09388 (2025)
68. Yue, X., Ni, Y., Zhang, K., Zheng, T., Liu, R., Zhang, G., Stevens, S., Jiang, D., Ren, W., Sun, Y., et al.: Mmmu: A massive multi-discipline multimodal understanding and reasoning benchmark for expert agi. In: CVPR (2024)
69. Zhang, H., Guo, Y., Kankanhalli, M.: Joint vision-language social bias removal for clip. In: CVPR (2025)
70. Zhang, Y., Chen, L., Zheng, G., Gao, Y., Zheng, R., Fu, J., Yin, Z., Jin, S., Qiao, Y., Huang, X., et al.: Spa-vl: A comprehensive safety preference alignment dataset for vision language models. In: CVPR (2025)
71. Zhao, J., Wang, T., Yatskar, M., Ordonez, V., Chang, K.W.: Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In: EMNLP (2017)
72. Zhao, S., Duan, R., Wang, F., Chen, C., Kang, C., Ruan, S., Tao, J., Chen, Y., Xue, H., Wei, X.: Jailbreaking multimodal large language models via shuffle inconsistency. In: ICCV (2025)
73. Zhao, Z., Yamasaki, T.: Bias beyond demographics: Probing decision boundaries in black-box lvlms via counterfactual vqa. arXiv preprint arXiv:2508.03079 (2025)
74. Zhu, J., Wang, W., Chen, Z., Liu, Z., Ye, S., Gu, L., Tian, H., Duan, Y., Su, W., Shao, J., et al.: Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. arXiv preprint arXiv:2504.10479 (2025)
75. Zong, Y., Bohdal, O., Yu, T., Yang, Y., Hospedales, T.: Safety fine-tuning at (almost) no cost: A baseline for vision large language models. In: ICML (2024)