Evaluating Fairness in Large Vision-Language Models Across Diverse Demographic Attributes and Prompts

Anonymous ACL submission

Abstract

Large vision-language models (LVLMs) have recently achieved significant progress, demonstrating strong capabilities in open-world visual understanding. However, it is not yet clear how LVLMs address demographic biases in real life, especially the disparities across attributes such as gender, skin tone, and age. In this paper, we empirically investigate *visual fairness* in several mainstream LVLMs and audit their performance disparities across sensitive demographic attributes, based on public fairness benchmark datasets (e.g., FACET). To disclose the visual bias in LVLMs, we design a fairness evaluation framework with direct questions and single-choice question-instructed prompts on visual question-answering/classification tasks. The zero-shot prompting results indicate that, despite enhancements in visual understanding, both open-source and closed-source LVLMs exhibit prevalent fairness issues across different instruct prompts and demographic attributes.

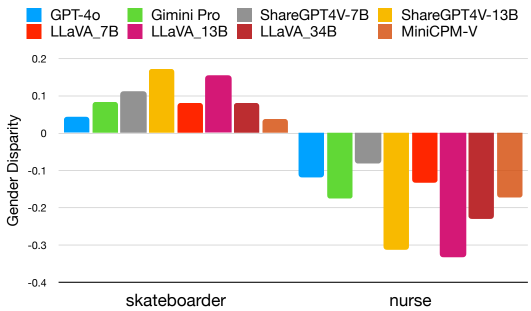

Figure 1: Gender disparity in person classes *[skateboarder, nurse]* across LVLMs in our experiments. Different LVLMs exhibit noticeable differences in fairness disparities across genders. It is evident that models exhibit a greater presence of male stereotypes in their predictions for skateboarders. Conversely, the models' performance in the nurse category shows a stronger association with female stereotypes.

1 Introduction

Large vision-language models (LVLMs) have successfully encoded images and text into a shared latent space, enabling a better visual reasoning (Radford et al., 2021; Jia et al., 2021). Pre-trained LVLMs can accurately interpret images and extract semantics by meticulously designing natural language instructions (also known as "prompts"), providing additional information for traditional vision tasks such as classification (Petryk et al., 2022; Abdelfattah et al., 2023), segmentation (Wang et al., 2022; He et al., 2023), and visual question answering (Zhu et al., 2023; Zhang et al., 2023). Although many studies and models have achieved remarkable results (OpenAI, 2023; Anil et al., 2023), there is a knowledge gap in the literature regarding the fairness evaluation of recent large models. Most existing works focus on improving the accuracy and efficiency of LVLMs (Liu et al., 2023a, 2024; Chen et al., 2023; Yu et al., 2024), with limited attention given to their performance across different demographic groups. This oversight is critical as it can lead to biased outcomes, potentially perpetuating stereotypes (Parraga et al., 2023), as illustrated in Figure 1 from our experiments. Moreover, existing studies (Chen et al., 2024; Han et al., 2023) have not adequately addressed the need for fairness evaluation specifically designed for the contemporary large model settings. It is essential to systematically study the impact of various demographic attributes on LVLMs performance.

In this study, we empirically provide a detailed evaluation of LVLMs from a fairness perspective. We propose a novel evaluation framework that employs direct questions and single-choice question-instructed prompts on visual question answering/classification tasks based on the FACET benchmark (Gustafson et al., 2023). The proposed framework analyzes the models' ability to understand and interpret images accurately while assessing any inherent biases related to visual clues such as gender, skin tone, and age. We summarize

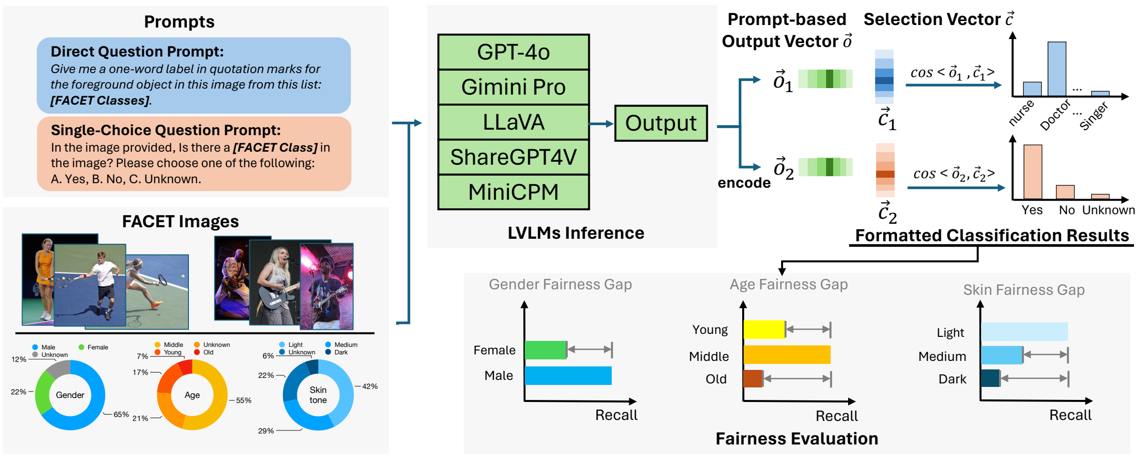

Figure 2: Proposed LVLMs fairness evaluation framework, showing the flow from FACET image collection to performance evaluation, highlighting the use of different types of instruct prompts and the detailed analysis of the model's responses.

Size	21,560 images, 21,560 people
Evaluation Annotations	52 - person related class
Demographic Attributes	
Gender	Male (14,110), Female (4,784), Unknown (2,666)
Age	Young (3,666), Middle (11,791), Old (1,513), Unknown (4,590)
Skin tone	Light (9154), Medium (6325), Dark (1314), Unknown (4767)

Table 1: Statistics of proposed evaluation dataset.

the contribution of this work in two folds: 1) We proposed a novel evaluation framework to investigate visual fairness issues in LVLMs, utilizing a fairness benchmark and meticulously designed instruct prompts. 2) Our extensive experimental results demonstrate that both open-source and closed-source LVLMs exhibit fairness issues across different instruct prompts and demographic attributes.

2 LVLMs Fairness Evaluation

2.1 Datasets Construction

To evaluate demographic bias in LVLMs based on attributes such as age, gender, and skin tone, we selected only images containing a single person from the FACET (Gustafson et al., 2023), a human-annotated fairness benchmark. Each image is annotated with demographic attributes, allowing us to systematically assess models' performance and identify visual fairness across different ages, genders, and skin tones in LVLMs. The statistics of our FACET dataset are shown in Table 1.

2.2 Evaluation Framework

Our LVLMs evaluation framework employs a variety of instruct prompts and a wide range of images in different scenarios. This framework is designed to assess the model's ability to understand individuals in images during prediction and classification tasks. By analyzing the results, we evaluate the model's performance across different demographic attributes, providing insights into its fairness and potential biases. Figure 2 illustrates our proposed LVLMs fairness evaluation framework.

Prompts Recent studies have shown that prompting methods are highly effective for evaluating LVLMs and LLMs (Liu et al., 2023b; Wang et al., 2024; Li et al., 2023b). Building on these studies, we designed specific prompts for LVLMs with different objectives by converting knowledge facts into a question-answering format. In our evaluation experiments, we use diverse instruct prompts tailored to extract person-related classes (e.g., soldier, nurse) from the images. **Direct Question Prompts** ask straightforward questions to gather specific information from the model, allowing for detailed responses. This approach provides in-depth insights into the model's understanding and generates rich, descriptive answers, making it ideal for exploratory analysis and assessing the model's comprehension. **Single-Choice Question Prompts** present a specific question with a set of predefined answers from which the model must choose, ensuring consistent and comparable responses. This method is effective for quantifying the model's accuracy and systematically detecting biases. More details of Prompts can be found in Appendix A.1.

LVLMs Inference and Formatting Results During model inference, the model generates predictions based on the instructed prompts and the content of the image. For direct question prompts, the model directly predicts the class label of the person in the image. For single-choice question prompts, the model answers based on the prompt about the person's class and the attributes in the image, providing the most probable prediction of yes, no, or unknown. Due to the LVLMs' unexpected output format issues (such as format errors or additional explanations), an encoder function encodes these raw labels as $\vec{o_1}$ and $\vec{o_2}$ and the selected respective labels $\vec{c_1}$ and $\vec{c_2}$ based on different prompt. The encoder finds the closest match using the cosine similarity function $cos{<}\vec{o},\vec{c}{>}$ (Li et al., 2023a). This method allows us to measure the likeness between the LVLMs' generated labels and the available dataset labels. More details of encoder functions can be found in Appendix A.2.

Evaluation Metrics We evaluate the performance of the models through two main aspects. First, we assess the model's understanding of the images by examining the accuracy of the model's predictions for the class of the person depicted in the image. Second, we perform a quantitative analysis of the impact of demographic attributes on the model's predictions. More details of demographic attributes illustrate in Appendix A.3.

We following the same fairness evaluation metric in FACET benchmark (Gustafson et al., 2023). Given a model f, the instruct prompt p, a set person class C, the demographic attribute l and a set of images I_l^C, we evaluate the model prediction accuracy based on recall, which compute by $R_l = recall(f(l, I_l^C, C))$. The value of R_l ranges between 0 and 1, with higher values indicating more accurate model predictions. We evaluate the model fairness by disparity between demographic attribute, which compute as $D_{l_1 - l_2} = R_{l_1} - R_{l_2} = recall(f(l_1, I_{l_1}^C, C)) - recall(f(l_2, I_{l_2}^C, C))$. When $D > 0$, the model exhibits a preference for l_1 within class c. Conversely, when $D < 0$, the model shows a preference for l_2 within class c. A disparity value of 0 indicates a perfectly fair model, demonstrating equal performance across all images within class c regardless of the demographic attributes $l1$ and $l2$.

3 Experiments

3.1 Experimental Settings

We evaluate various LVLMs, including both closed-source and open-source models, under a zero-shot setting to assess their ability to generate accurate answers without fine-tuning. Customized prompts from our framework are used for each model evaluation based on the specific model inference setting. All experiments are conducted using NVIDIA A100 GPUs.

Evaluation Models We utilize CLIP (Radford et al., 2021) and ViT (Dosovitskiy et al., 2021) as our baseline models, which align visual and textual representations to enable zero-shot learning across diverse vision tasks. We report the classification results for the person class only due to model evaluation limitations. For closed-source LVLMs, we select GPT-4o (OpenAI, 2023) and Gimini 1.5 Pro (Anil et al., 2023). For open-source LVLMs, we include LLaVa-1.5 (7B and 13B parameters versions) (Liu et al., 2023a), LLaVa-1.6 (34B version) (Liu et al., 2024), ShareGPT4V (7B and 13B versions) (Chen et al., 2023), and MiniCPM-V (8B version) (Yu et al., 2024). These LVLMs have demonstrated significant vision understanding abilities across various benchmark datasets.

3.2 Results and Analysis

In Table 2, we present the overall evaluation results of recall and disparity for each demographic group (gender, skin tone and age) from each model, based on images of 13 selected person classes. Detailed results for each class and each model will be provided in the Appendix A.4. Despite improvements in recall accuracy, nearly all LVLMs exhibit fairness issues across gender, skin tone, and age, leading to biased outcomes and perpetuating existing inequalities.

Models Except for the 7B-based models, other LVLMs show significant improvements in recall performance over traditional CLIP and ViT models, indicating enhanced image understanding and increasing accuracy with more model parameters. However, LVLMs have not shown significant improvements in fairness metrics, with some performing worse than the baselines. Closed-source LVLMs do not have absolute superiority over open-source LVLMs in recall performance and fairness metrics. For instance, GPT-4 and Gimini 1.5 Pro often respond with "unknown" to sensitive questions when information is insufficient, unlike open-source models, which tend to provide vague answers. It reveals that even the most accurate models can still perform inconsistently across different demographic groups.

Demographic Groups In evaluating gender-based performance, LVLMs fairness assessments reveal differing disparities depending on the prompt type. Direct question prompts tend to elicit more stereotypically female attributes, while single-

Model	Direct Question Prompt			Single-Choice Question Prompt		
	R_{Male}	R_{Female}	$D_{Male-Female}$	R_{Male}	R_{Female}	$D_{Male-Female}$
CLIP	0.5739	0.5482	0.0257	N/A	N/A	N/A
ViT	0.4957	0.5163	-0.0206	N/A	N/A	N/A
GPT-4o	0.7124	0.7386	-0.0262	0.8055	0.6970	0.1086
Gimini 1.5 Pro	0.7372	0.7584	-0.0212	0.8260	0.7753	0.0507
LLaVA-1.5 (7B)	0.5035	0.5151	-0.0115	0.9401	0.9120	0.0280
LLaVA-1.5 (13B)	0.6258	0.6741	-0.0483	0.8218	0.7410	0.0808
ShareGPT4V (7B)	0.5509	0.5976	-0.0467	0.9178	0.8988	0.0190
ShareGPT4V(13B)	0.6674	0.7072	-0.0399	0.7770	0.7090	0.0680
MiniCPM-V (8B)	0.6676	0.6669	0.0008	0.8561	0.8331	0.0229
LLaVA-1.6 (34B)	0.6558	0.6970	-0.0411	0.8393	0.8072	0.0321

(a) Performance on Demographic Gender

Model	Direct Question Prompt				Single-Choice Question Prompt			
	R_{Light}	R_{Medium}	R_{Dark}	$D_{Light-Dark}$	R_{Light}	R_{Medium}	R_{Dark}	$D_{Light-Dark}$
CLIP	0.6070	0.5436	0.4369	0.1701	N/A	N/A	N/A	N/A
ViT	0.5429	0.4662	0.4523	0.0906	N/A	N/A	N/A	N/A
GPT-4o	0.7473	0.7112	0.6185	0.1288	0.7798	0.7745	0.7692	0.0105
Gimini 1.5 Pro	0.7644	0.7319	0.6492	0.1151	0.8122	0.8093	0.8215	-0.0093
LLaVA-1.5 (7B)	0.5512	0.4759	0.3754	0.1758	0.9371	0.9244	0.9262	0.0110
LLaVA-1.5 (13B)	0.6919	0.6069	0.5231	0.1688	0.8043	0.7745	0.8092	-0.0049
ShareGPT4V (7B)	0.6141	0.5442	0.3815	0.2325	0.9172	0.9062	0.9015	0.0156
ShareGPT4V (13B)	0.7227	0.6508	0.5631	0.1597	0.7623	0.7459	0.7385	0.0238
MiniCPM-V (8B)	0.7044	0.6569	0.5292	0.1752	0.8639	0.8355	0.8215	0.0423
LLaVA-1.6 (34B)	0.7123	0.6362	0.5292	0.1831	0.8422	0.8202	0.8185	0.0238

(b) Performance on Demographic Skin Tone Groups

Model	Direct Question Prompt				Single-Choice Question Prompt			
	R_{Young}	R_{Middle}	R_{Old}	$D_{Young-Old}$	R_{Young}	R_{Middle}	R_{Old}	$D_{Young-Old}$
CLIP	0.6267	0.5587	0.4722	0.1545	N/A	N/A	N/A	N/A
ViT	0.5949	0.4986	0.3355	0.2594	N/A	N/A	N/A	N/A
GPT-4o	0.7753	0.7087	0.6987	0.0766	0.7745	0.7822	0.7415	0.0330
Gimini 1.5 Pro	0.8017	0.7316	0.6944	0.1073	0.8258	0.8216	0.7650	0.0609
LLaVA-1.5 (7B)	0.5723	0.5097	0.3932	0.1792	0.9479	0.9326	0.9145	0.0334
LLaVA-1.5 (13B)	0.7333	0.6321	0.5192	0.2141	0.8009	0.8092	0.7372	0.0638
ShareGPT4V (7B)	0.6439	0.5491	0.5085	0.1353	0.9269	0.9180	0.8761	0.0508
ShareGPT4V (13B)	0.7566	0.6674	0.6303	0.1263	0.7784	0.7638	0.7051	0.0733
MiniCPM-V (8B)	0.7286	0.6582	0.6090	0.1196	0.8538	0.8591	0.8162	0.0376
LLaVA-1.6 (34B)	0.7675	0.6496	0.6368	0.1307	0.8546	0.8417	0.7735	0.0811

(c) Performance on Demographic Age Groups

Table 2: Overall evaluation of model performance in recall and disparity for each demographic group (Gender, Skin Tone, and Age) based on images from selected person classes. Closed-source LVLMs highlighted in light gray.

choice prompts lean towards male attributes. For the demographic attribute of skin tone, the performance under the direct question prompt shows a clear preference for lighter skin tones over darker ones. This bias is also evident in the age group evaluation, where the direct question prompt demonstrates a tendency to favor younger individuals over older ones.

Prompts Based on various prompts, single-choice question prompt generally achieve higher recall performance than direct question prompt for the same images across all demographic groups. This trend is especially pronounced in open-source LVLMs, which show a significant performance gap. Conversely, closed-source LVLMs exhibit smaller gaps and more consistent outputs. In fairness evaluations, single-choice question prompt consistently yield lower disparity scores.

4 Conclusion and Future Work

In this paper, we proposed the novel visual fairness evaluation framework for investigating demographic bias in LVLMs. The experimental results demonstrated significant fairness gap across gender, skin tone, and age in both open-source and closed-source LVLMs. In future work, we aim to fine-tune LVLMs by incorporating fairness constraints and bias mitigation techniques to reduce disparities.

5 Limitations

Our study provides a novel evaluation of LVLMs from a fairness perspective, it still has several limitations. 1) The dataset may not fully capture all real-world demographic attributes, and the design of instruct prompts may not cover all dimensions of bias. 2) The model output can vary across different versions and configurations of models, particularly with close-source LVLMs that lack transparency. 3) Our evaluation framework might not reflect the evolving nature of biases, and the focus on gender, skin tone, and age may not cover other critical demographic factors. 4) The high computational resources required for this framework may limit its applicability. Addressing these limitations will be crucial for better evaluating fairness in LVLMs.

References

Rabab Abdelfattah, Qing Guo, Xiaoguang Li, Xiaofeng Wang, and Song Wang. 2023. CDUL: clip-driven unsupervised learning for multi-label image classification. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 1348–1357. IEEE.

Rohan Anil, Sebastian Borgeaud, Yonghui Wu, and et al. 2023. Gemini: A family of highly capable multimodal models. *CoRR*, abs/2312.11805.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Conghui He, Jiaqi Wang, Feng Zhao, and Dahua Lin. 2023. Sharegpt4v: Improving large multi-modal models with better captions. *CoRR*, abs/2311.12793.

Lin Chen, Jinsong Li, Xiaoyi Dong, Pan Zhang, Yuhang Zang, Zehui Chen, Haodong Duan, Jiaqi Wang, Yu Qiao, Dahua Lin, and Feng Zhao. 2024. Are we on the right way for evaluating large vision-language models? *CoRR*, abs/2403.20330.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

Laura Gustafson, Chloé Rolland, Nikhila Ravi, Quentin Duval, Aaron Adcock, Cheng-Yang Fu, Melissa Hall, and Candace Ross. 2023. FACET: fairness in computer vision evaluation benchmark. In *IEEE/CVF International Conference on Computer Vision, ICCV 2023, Paris, France, October 1-6, 2023*, pages 20313–20325. IEEE.

Xiaotian Han, Jianfeng Chi, Yu Chen, Qifan Wang, Han Zhao, Na Zou, and Xia Hu. 2023. Ffb: A fair fairness benchmark for in-processing group fairness methods. *Preprint*, arXiv:2306.09468.

Wenbin He, Suphanut Jamonnak, Liang Gou, and Liu Ren. 2023. CLIP-S4: language-guided self-supervised semantic segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 11207–11216. IEEE.

Courtney M. Heldreth, Ellis P. Monk, Alan T. Clark, Candice Schumann, Xango Eyee, and Susanna Ricco. 2024. Which skin tone measures are the most inclusive? an investigation of skin tone measures for artificial intelligence. *ACM J. Responsib. Comput.*, 1(1).

Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc V. Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. 2021. Scaling up visual and vision-language representation learning with noisy text supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 4904–4916. PMLR.

Lei Li, Jingjing Xu, Qingxiu Dong, Ce Zheng, Xu Sun, Lingpeng Kong, and Qi Liu. 2023a. Can language models understand physical concepts? In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, EMNLP 2023, Singapore, December 6-10, 2023*, pages 11843–11861. Association for Computational Linguistics.

Yaowei Li, Ruijie Quan, Linchao Zhu, and Yi Yang. 2023b. Efficient multimodal fusion via interactive prompting. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2023, Vancouver, BC, Canada, June 17-24, 2023*, pages 2604–2613. IEEE.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2023a. Improved baselines with visual instruction tuning. *CoRR*, abs/2310.03744.

Haotian Liu, Chunyuan Li, Yuheng Li, Bo Li, Yuanhan Zhang, Sheng Shen, and Yong Jae Lee. 2024. Llava-next: Improved reasoning, ocr, and world knowledge.

Xiaoxia Liu, Jingyi Wang, Jun Sun, Xiaohan Yuan, Guoliang Dong, Peng Di, Wenhai Wang, and Dongxia Wang. 2023b. Prompting frameworks for large language models: A survey. *CoRR*, abs/2311.12785.

OpenAI. 2023. Gpt-4v(ision) system card.

Otavio Parraga, Martin D. More, Christian M. Oliveira, Nathan S. Gavenski, Lucas S. Kupssinskü, Adilson Medronha, Luis V. Moura, Gabriel S. Simões, and Rodrigo C. Barros. 2023. Fairness in deep learning: A survey on vision and language research. *ACM Comput. Surv.* Just Accepted.

Suzanne Petryk, Lisa Dunlap, Keyan Nasseri, Joseph Gonzalez, Trevor Darrell, and Anna Rohrbach. 2022. On guiding visual attention with language specification. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 18071–18081. IEEE.

Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, Gretchen Krueger, and Ilya Sutskever. 2021. Learning transferable visual models from natural language supervision. In *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Yuan Wang, Xuyang Wu, Hsin-Tai Wu, Zhiqiang Tao, and Yi Fang. 2024. Do large language models rank fairly? an empirical study on the fairness of llms as rankers. *CoRR*, abs/2404.03192.

Zhaoqing Wang, Yu Lu, Qiang Li, Xunqiang Tao, Yandong Guo, Mingming Gong, and Tongliang Liu. 2022. CRIS: clip-driven referring image segmentation. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pages 11676–11685. IEEE.

Tianyu Yu, Haoye Zhang, Yuan Yao, Yunkai Dang, Da Chen, Xiaoman Lu, Ganqu Cui, Taiwen He, Zhiyuan Liu, Tat-Seng Chua, and Maosong Sun. 2024. Rlaif-v: Aligning mllms through open-source ai feedback for super gpt-4v trustworthiness. *arXiv preprint arXiv:2405.17220*.

Pan Zhang, Xiaoyi Dong, Bin Wang, Yuhang Cao, Chao Xu, Linke Ouyang, Zhiyuan Zhao, Shuangrui Ding, Songyang Zhang, Haodong Duan, Wenwei Zhang, Hang Yan, Xinyue Zhang, Wei Li, Jingwen Li, Kai Chen, Conghui He, Xingcheng Zhang, Yu Qiao, Dahua Lin, and Jiaqi Wang. 2023. Internlm-xcomposer: A vision-language large model for advanced text-image comprehension and composition. *CoRR*, abs/2309.15112.

Deyao Zhu, Jun Chen, Xiaoqian Shen, Xiang Li, and Mohamed Elhoseiny. 2023. Minigpt-4: Enhancing vision-language understanding with advanced large language models. *CoRR*, abs/2304.10592.

A Appendix

A.1 Prompts

Table 3 illustrates the direct questions and single-choice question-instructed prompts utilized in our LVLMs fairnesss evaluation framework.

A.2 Encode Functions

In this study, we utilized two different text encoder methods: the CLIP text encoder and the T5 text encoder. These encoders were employed to enhance the matching between the outputs from LVLMs and the selected class labels. We used the pre-trained parameters of both models to leverage their robust capabilities.

A.3 Demographic Attributes

For gender presentation, we aim to investigate whether the model's predictions exhibit more stereotypically male attributes or more stereotypically female attributes. For skin tone, we categorize into three distinct groups based on The Monk Skin Tone Scale (Heldreth et al., 2024): light (Monk points 1-3), medium (Monk points 4-6), and dark (Monk points 7-10) (Heldreth et al., 2024). For age, we classify into three perceived age groups: younger (under 25 years old), middle-aged (25-65 years old), and older (over 65 years old).

A.4 Class-level Evaluation Results

To provide a deeper understanding, detailed results for each individual class and each model, this supplementary information allows for an in-depth analysis of how each model performs across various person classes and demographic groups, ensuring a robust evaluation of both accuracy and fairness.

Name	Content
Direct Question Prompt	
p1	Give me a one-word label in quotation marks for the foreground object in this image from this list: *[FACET Classes]*
Single-Choice Question Prompt	
p2	In the image provided, Is there a *[FACET class]* in the image? Please choose one of the following: A. Yes, B. No, C. Unknown.
p3	In the image provided, Do you think the person in the image is a *[FACET class]* ? Please choose one of the following: A. Yes, B. No, C. Unknown.

Table 3: Different instruct prompts utilized in LVLMs fairness evaluation framework.

Model	gardener	craftsman	laborer	skateboarder	prayer	guitarist	singer	dancer	retailer	nurse	student	gymnast	horseman
GPT-4o	-0.0040	0.0041	0.0338	0.0366	-0.0178	0.1676	-0.0739	-0.1434	-0.1721	-0.3425	-0.0251	0.0834	-0.0302
Gimini 1.5 Pro	0.0362	-0.0075	-0.0170	0.0508	-0.0227	0.1377	-0.0659	-0.0490	-0.1770	-0.3707	-0.0995	-0.0387	-0.0346
LLaVA-1.5 (7B)	-0.0407	-0.1461	0.0097	0.1052	-0.1054	0.1573	-0.1024	-0.1282	-0.1187	-0.0678	0.0184	0.0275	-0.1711
LLaVA-1.5 (13B)	-0.0087	-0.0874	0.0644	0.0920	0.0520	0.0647	-0.1463	-0.3089	-0.1862	-0.2208	-0.1111	-0.0616	-0.0578
ShareGPT4V (7B)	-0.0841	-0.3031	0.0289	0.0878	0.0436	0.0644	-0.1433	-0.1305	-0.1951	-0.0615	-0.0966	-0.0750	-0.0894
ShareGPT4V(13B)	-0.0154	0.0717	0.0862	0.0741	-0.0030	0.0748	-0.1049	-0.2413	-0.2410	-0.3264	-0.0638	-0.0035	-0.0692
MiniCPM-V (8B)	0.0371	-0.0151	0.0086	0.0815	0.0032	0.0971	-0.0848	-0.1305	0.0184	-0.2443	-0.1990	0.0095	-0.0368
LLaVA-1.6 (34B)	-0.0680	0.0130	-0.0189	0.0284	0.0253	0.3036	-0.0565	-0.1783	-0.1944	-0.1881	-0.0174	-0.0352	-0.0420

(a) Fairness Performance Disparity between Male and Female of Selected Classes Based on Direct Question Prompts.

Model	gardener	craftsman	laborer	skateboarder	prayer	guitarist	singer	dancer	retailer	nurse	student	gymnast	horseman
GPT-4o	0.1516	0.0543	0.1407	0.0443	-0.0237	0.1398	0.0104	-0.0589	-0.0777	-0.1201	0.0068	-0.1061	0.0451
Gimini 1.5 Pro	0.1279	0.0919	0.1105	0.0832	-0.0104	0.1229	-0.0209	-0.0495	-0.0542	-0.1747	-0.0271	-0.1092	0.0217
LLaVA-1.5 (7B)	0.1039	0.1730	0.0942	0.0805	0.0471	0.0589	0.0042	-0.0501	-0.0514	-0.1320	-0.0271	-0.0493	0.0280
LLaVA-1.5 (13B)	0.0788	0.2326	0.2097	0.1537	0.0001	0.2148	-0.0212	-0.2523	-0.1475	-0.3327	-0.0464	-0.0887	0.0457
ShareGPT4V (7B)	0.0181	0.0457	0.0354	0.1117	0.0065	0.0689	0.0062	-0.0967	-0.0766	-0.0828	-0.0937	-0.0554	0.0759
ShareGPT4V(13B)	0.0941	0.1772	0.2040	0.1724	-0.0046	0.1050	-0.0429	-0.2914	-0.1418	-0.3136	-0.0386	-0.1041	0.1363
MiniCPM-V (8B)	0.0833	0.0481	0.1043	0.0374	-0.0369	0.0748	-0.0033	-0.1002	-0.1082	-0.1722	-0.1285	-0.1211	0.0122
LLaVA-1.6 (34B)	0.1480	0.0581	0.1514	0.0810	-0.0334	0.1092	-0.0053	-0.1387	-0.1720	-0.2295	-0.0232	-0.1122	0.0128

(b) Fairness Performance Disparity between Male and Female of Selected Classes Based on Single-Choice Question Prompts.

Table 4: Fairness Performance Disparity between Male and Female of Selected Classes. Closed-source LVLMs highlighted in light gray.

Model	gardener	craftsman	laborer	skateboarder	prayer	guitarist	singer	dancer	retailer	nurse	student	gymnast	horseman
GPT-4o	-0.0901	-0.0520	-0.0278	0.0157	0.0100	0.0417	0.0683	0.2224	-0.1343	0.1614	-0.0123	-0.1191	-0.0437
Gimini 1.5 Pro	0.1409	-0.0386	-0.0510	0.0611	0.0150	0.0837	-0.0059	0.1413	0.0537	0.1228	0.1520	0.0977	-0.0786
LLaVA-1.5 (7B)	0.0959	-0.1528	-0.0122	-0.0208	-0.3509	0.1554	0.1669	0.1275	0.0940	-0.1263	-0.0539	0.3182	0.2860
LLaVA-1.5 (13B)	0.1229	-0.0883	-0.0575	0.0223	-0.1424	0.0652	0.0012	0.1945	-0.1224	-0.0632	0.1593	0.1527	-0.0873
ShareGPT4V (7B)	0.0882	-0.0712	-0.0077	-0.0009	0.0341	0.0757	0.2723	0.2671	-0.1776	-0.0386	0.2598	0.1645	-0.1223
ShareGPT4V (13B)	-0.1351	-0.1240	-0.0169	0.0223	-0.1559	0.1039	0.0919	0.3843	-0.1224	0.0246	-0.0172	0.1786	-0.0655
MiniCPM-V (8B)	0.0869	-0.0556	0.0145	0.0223	0.0105	0.1708	0.0781	0.1863	-0.1582	0.0842	-0.1887	0.1027	0.2020
LLaVA-1.6 (34B)	0.0431	-0.0470	-0.0467	-0.0066	0.0627	0.0908	0.0592	0.0464	-0.1597	0.0456	0.0539	0.1268	-0.0742

(a) Fairness Performance Disparity between Light and Dark of Selected Classes Based on Direct Question Prompts.

Model	gardener	craftsman	laborer	skateboarder	prayer	guitarist	singer	dancer	retailer	nurse	student	gymnast	horseman
GPT-4o	-0.1203	-0.0450	-0.0928	0.0015	-0.1704	0.0999	0.1074	0.0610	0.0985	-0.0281	0.2255	0.2295	0.1496
Gimini 1.5 Pro	-0.2259	-0.0560	-0.1561	0.0569	-0.2496	0.1328	0.1023	0.0159	0.0582	-0.0211	0.2770	0.1486	0.1801
LLaVA-1.5 (7B)	-0.0727	-0.0756	-0.0824	0.0379	-0.1048	0.0427	0.0283	0.0520	0.1881	0.1930	0.1716	-0.0400	0.2369
LLaVA-1.5 (13B)	-0.0914	-0.0731	-0.1455	0.0313	-0.1549	0.1305	0.0319	0.2379	0.0597	0.1579	0.0539	0.2305	0.1714
ShareGPT4V (7B)	0.0257	-0.0134	-0.0721	0.0644	-0.2837	0.0894	0.0521	0.1550	0.0731	0.0842	0.3358	0.1018	-0.0480
ShareGPT4V (13B)	-0.1281	-0.0132	-0.1662	-0.0084	-0.0446	0.0757	0.0657	0.4212	0.1134	0.1333	0.1201	0.2305	0.1059
MiniCPM-V (8B)	-0.1178	-0.0536	-0.0961	0.0801	0.0566	0.1627	0.0667	0.1408	0.0060	0.2456	0.2181	0.2995	0.2107
LLaVA-1.6 (34B)	-0.1358	-0.0523	-0.1049	0.0512	-0.2737	0.0918	0.0823	0.0674	0.0313	0.1754	0.2843	0.2595	0.2282

(b) Fairness Performance Disparity between Light and Dark of Selected Classes Based on Single-Choice Question Prompts.

Table 5: Fairness Performance Disparity between Light and Dark of Selected Classes. Closed-source LVLMs highlighted in light gray.

Model	gardener	craftsman	laborer	skateboarder	prayer	guitarist	singer	dancer	retailer	nurse	student	gymnast	horseman
GPT-4o	0.0109	-0.1648	-0.1061	0.9522	-0.0008	-0.0374	0.1421	-0.2893	0.3783	0.0791	0.7963	-0.2116	0.0684
Gimini 1.5 Pro	-0.0855	-0.1878	0.0198	0.9522	0.0403	-0.0900	0.2057	0.0269	0.2204	-0.0128	0.8889	0.3519	0.1263
LLaVA-1.5 (7B)	-0.1302	-0.1082	0.0105	0.9261	0.0880	-0.0097	0.0699	0.1198	0.0801	-0.0299	0.1852	0.4762	0.2895
LLaVA-1.5 (13B)	0.1043	-0.0048	0.0350	0.9783	-0.1077	-0.0510	0.1097	-0.0372	0.2921	0.1859	0.7222	0.8942	0.1158
ShareGPT4V (7B)	0.0109	-0.1025	0.0233	0.9478	-0.0428	-0.0474	0.1877	-0.1136	0.0656	0.0043	0.3889	0.7672	0.1421
ShareGPT4V (13B)	0.0825	-0.1662	-0.0186	0.9826	-0.0033	-0.0510	0.2371	-0.1302	0.3005	-0.0321	0.5741	0.3042	0.1474
MiniCPM-V (8B)	-0.0443	-0.1632	-0.0839	0.9696	-0.0962	-0.0751	0.2475	0.0950	0.1320	0.0021	0.7037	0.8519	0.0368
LLaVA-1.6 (34B)	-0.0105	-0.1761	-0.0478	0.9957	-0.1480	-0.1735	0.1001	-0.0888	0.1434	0.1432	0.8148	-0.0582	0.1263

(a) Fairness Performance Disparity between Young and Old of Selected Classes Based on Direct Question Prompts.

Model	gardener	craftsman	laborer	skateboarder	prayer	guitarist	singer	dancer	retailer	nurse	student	gymnast	horseman
GPT-4o	-0.0975	-0.0300	-0.1282	0.9043	0.1530	-0.0141	0.0729	-0.0558	0.0244	0.1197	0.7407	0.3148	0.1632
Gimini 1.5 Pro	-0.2644	-0.1062	0.0058	0.8957	0.1118	-0.0346	0.0023	-0.1818	-0.0183	-0.1667	0.8889	0.8413	0.1842
LLaVA-1.5 (7B)	-0.1894	0.0418	-0.0023	0.9652	-0.0740	-0.0241	0.0185	0.2087	-0.0008	0.0726	0.9074	0.4894	0.1474
LLaVA-1.5 (13B)	-0.2322	-0.0889	0.1014	0.9478	0.0979	-0.0049	0.0580	0.1116	0.1793	0.2094	0.7407	0.7460	0.1632
ShareGPT4V (7B)	-0.1913	-0.0445	-0.0163	0.9739	0.0617	-0.0241	0.0608	0.1756	-0.0008	0.0150	0.9444	0.4471	-0.1053
ShareGPT4V (13B)	-0.2142	-0.0329	-0.0455	0.9348	0.1242	0.0044	0.0499	-0.0393	0.1076	0.2671	0.7593	0.7672	0.0474
MiniCPM-V (8B)	-0.2753	-0.0387	-0.0653	0.9130	-0.1349	-0.0418	0.0367	-0.1901	-0.1060	-0.1004	0.8889	0.8730	0.2368
LLaVA-1.6 (34B)	-0.2573	-0.0344	-0.0490	0.9652	0.1234	0.0072	0.1056	0.0764	-0.1152	-0.0470	0.7037	0.8624	0.1684

(b) Fairness Performance Disparity between Young and Old of Selected Classes Based on Single-Choice Question Prompts.

Table 6: Fairness Performance Disparity between Young and Old of Selected Classes. Closed-source LVLMs highlighted in light gray.