Assessing Multilingual Fairness in Pre-trained Multimodal Representations

Jialu Wang and **Yang Liu** and **Xin Eric Wang**
Department of Computer Science and Engineering
University of California, Santa Cruz
{faldict,yangliu,xwang366}@ucsc.edu

Abstract

Recently pre-trained multimodal models, such as CLIP (Radford et al., 2021), have shown exceptional capabilities towards connecting images and natural language. The textual representations in English can be desirably transferred to multilingualism and support downstream multimodal tasks for different languages. Nevertheless, the principle of multilingual fairness is rarely scrutinized: do multilingual multimodal models treat languages equally? Are their performances biased towards particular languages? To answer these questions, we view language as the fairness recipient and introduce two new fairness notions, multilingual individual fairness and multilingual group fairness, for pre-trained multimodal models. Multilingual individual fairness requires that text snippets expressing similar semantics in different languages connect similarly to images, while multilingual group fairness requires equalized predictive performance across languages. We characterize the extent to which pre-trained multilingual vision-and-language representations are individually fair across languages. However, extensive experiments demonstrate that multilingual representations do not satisfy group fairness: (1) there is a severe multilingual accuracy disparity issue; (2) the errors exhibit biases across languages conditioning the group of people in the images, including race, gender and age.

1 Introduction

Recently pre-trained vision-and-language representations (Lu et al., 2019; Tan and Bansal, 2019; Su et al., 2020; Li et al., 2020a; Chen et al., 2020; Li et al., 2020b; Gan et al., 2020; Yu et al., 2021; Desai and Johnson, 2021; Radford et al., 2021; Cho et al., 2021) have received a surge of attention. Such pre-trained multimodal representations have shown great capabilities of bridging images and natural language on the downstream tasks, including image captioning (Laina et al., 2019), image retrieval (Vo et al., 2019), visual QA (Zhou et al., 2020), text-to-image generation (Ramesh et al., 2021), etc. While it is commonly recognized that the multimodal representations trained on English corpora can be generalized to multilingualism by cross-lingual alignment (Lample and Conneau, 2019; Conneau et al., 2020), recent studies criticize that the multilingual textual representations *do not* learn equally high-quality representations for all the languages (Wu and Dredze, 2020), especially for low-resource languages. Hu et al. (2020) emphasize the need for general-purpose representations to seek equal performance across all languages. However, there is still a lack of a nuanced understanding of how multilingual representations fare on vision-and-language benchmarks.

This paper provides a novel perspective for analyzing the principles of multilingual fairness in multimodal representations from two aspects. First, existing frameworks for measuring multilingual biases usually emulate text sources in different languages, which may have ambiguous meanings in varied contexts (González et al., 2020). In contrast, we leverage visual grounding as the anchor to bridge text in different languages—text snippets in different languages but with similar semantics should be equitably relevant to the same images. Second, we equate a language as an aggregated group of individuals (e.g., French as a group of French sentences) in the terminology of fairness. As Choudhury and Deshpande (2021) has pointed out, "*each language has a distinct identity, defined by its vocabulary, syntactic structure, its typological features, amount of available resources, and so on.*" The notions of fairness, such as individual fairness (Dwork et al., 2012) and group fairness (Zemel et al., 2013; Chouldechova, 2017; Hardt et al., 2016; Zhu et al., 2022), can be naturally adapted by comparing the multimodal model's treatment across languages.

Therefore, we introduce two fairness notions:

Findings of the Association for Computational Linguistics: ACL 2022, pages 2681 - 2695
May 22-27, 2022 ©2022 Association for Computational Linguistics

multilingual individual fairness presumes similar outcomes between similar language expressions grounding on the same images; *multilingual group fairness* postulates that multimodal models should induce similar predictive performance across different languages. These fairness notions are formalized to *compare the multimodal model's treatment of one language versus another* for either the individual target or the aggregated group.

Our contributions are as follows:

- We formally define the individual fairness and group fairness notions in the multilingual and multimodal setting (see Section 3 and Section 4).
- We theoretically investigate the extent to which pre-trained multilingual vision-and-language representations are individually fair. However, our negative result demonstrates that individual fairness does not suffice to prevent accuracy disparity at the group level (see Section 5.1).
- Extensive experimental results reveal the accuracy disparity across different languages. Our results also imply that the choice of visual representations affects the group fairness metrics (see Section 5.2).
- We further demonstrate the prevalence of group rate disparity when language is coupled with multi-dimensional groups associated with images, such as race, gender, and age (see Section 5.3). Our empirical exploration provides new directions for mitigating biases under the multilingual setting.

2 Background

Notation. Throughout the paper, we use the uppercase letter I to denote images and T to denote text. We use the superscript (L) in $T^{(L)}$ to represent the text is in language L. When we are jointly using $T^{(L)}$ and $T^{(L')}$ for two languages L and L', we often assume that they share the same semantic meanings. Lowercase letters $\boldsymbol{v}$ and $\boldsymbol{t}$ are used to denote the visual and textual representation vectors encoded by model M, respectively. To simplify the presentation, we use $S(\cdot,\cdot)$ to generally represent the similarity between images and text. Specifically, $S(I,T)$ refers to the similarity scores predicted by the model M between the image I and text T, while $S(\boldsymbol{v},\boldsymbol{t})$ refers to the cosine similarity between vectors $\boldsymbol{v}$ and $\boldsymbol{t}$.

2.1 Multilingual CLIP

Our work is established on the multimodal setting. The universal framework for matching images and text (Mogadala et al., 2021) is to encode them into representation vectors in a shared representation space, such that the distance between visual and textual vectors can measure the similarity between images and text. Throughout this paper, our analysis mainly focuses on CLIP (Contrastive Language-Image Pre-training Radford et al., 2021), a representative pre-trained multimodal representation model that achieves state-of-the-art performances on zero-shot transfer tasks.

CLIP is a multimodal model trained on large-scale images with natural language supervision collected from the internet. It comprises an image encoder and a text encoder that can embed images and text into visual and textual representation vectors. One desirable property is that the CLIP model takes the cosine similarity between image and text features to measure the log-odds of the corresponding image-text pairs, and is trained to maximize their similarity by a contrastive learning objective. In light of this capability, CLIP can predict the similarity, denoted by $S(I,T)$, between arbitrary images I and natural language text snippets T.

In order to adapt the flexible CLIP model to multilingualism, Multilingual CLIP (Carlsson and Ekgren, 2021) uses a pre-trained multilingual language model, such as M-BERT (Devlin et al., 2019), to take over the original text encoder in English, and fine-tune the textual representation vectors by cross-lingual alignment (Lample and Conneau, 2019; Conneau et al., 2020). In this setting, we use $S(I,T^{(L)})$ to represent the similarity between image I and text $T^{(L)}$ in language L. Though the empirical evaluations in this paper mainly focus on Multilingual CLIP, the experimental approaches we adopt to arrive at the observations can be generalized to other pre-trained multilingual vision-and-language representations.

2.2 Fairness Notions

The multilingual fairness notions developed in this work is inspired by multiple fairness definitions (Narayanan, 2018; Dwork et al., 2012) in the algorithmic fairness literature. We will briefly introduce these fairness notions in fair decision making and instantiate them in the domain of multilingual vision-and-language learning later.

Individual fairness, initiated by Dwork et al.

(2012), requires that individuals who are similar with respect to a task-specific similarity metric have similar decision outcomes.

Group fairness definitions seek to provide fairness guarantees based on group-level statistical constraints, in the sense that they are evaluated and enforced without reference to similarity measures. In the fairness literature, group fairness is commonly framed in terms of *protected groups* G, such as race, gender, and age. For instance, demographic parity (Zafar et al., 2017) requires that the outcomes are independent of the group membership, and equalized odds (Hardt et al., 2016) essentially requires equal true positive and false positive rates between different groups.

Principally, these fairness criteria are formulated by comparing the treatment of one individual or one group versus another. Our work will instantiate the standard fairness notions by viewing language as the recipient — we compare *how the treatment of one language differs from another*.

2.3 Fairness in NLP

Many recent works (Choudhury and Deshpande, 2021; Hu et al., 2020; Pires et al., 2019; González et al., 2020; Ross et al., 2021) scrutinize the ethical issues raised in multilingual settings, albeit with varying degrees of success. For instance, Zhao et al. (2020) quantifies the presence of representational biases in multilingual word embeddings by calculating the distance between targets corresponding to different sensitive attributes. Huang et al. (2020) evaluate group fairness violations among demographic groups on the task of hate speech detection, but do not explicitly regard language as unique group membership. Burns et al. (2020) studies the performance degradation when multimodal models are trained to support additional languages, and tries to address the multilingual accuracy disparity on the task of image-sentence retrieval. Our work complements the fairness discourse in multilingual NLP to the extent that we provide a novel perspective of studying multilingual fairness by viewing language as the recipient of fairness notions.

Our work is also closely relevant to prior studies on biases in vision-and-language tasks, including visual semantic role labeling (Zhao et al., 2017), image captioning (Burns et al., 2018; Tang et al., 2021), and image search (Wang et al., 2021). Notably, Srinivasan and Bisk (2021) investigates the gender bias associated with entities for pre-trained representations. Compared to these works, we focus on generic fairness measures for multimodal models and use visual grounding to bridge different languages.

3 Multilingual Individual Fairness

For an ideal multilingual vision-and-language model, text descriptions in different languages referring to similar semantic meanings should be equally similar or dissimilar to the same grounding images. We note that there are no language expressions that are perfectly identical to each other in real-world scenarios due to linguistic features. Nevertheless, at least in a normal vision-and-language task, multilingual models are desired to impose equal treatment to different languages. For instance, “this is a cat” (in English) and “das ist eine Katze” (in German) should be similarly related to an image of a cat in image-text retrieval. This intuition aligns with individual fairness in a multilingual manner. In this section, we investigate to what degree multilingual representations are individually fair.

Individual fairness requires that similar people should be treated similarly (Dwork et al., 2012). In our multilingual setting, we require that the text snippets expressing similar semantics in different languages should be similarly related to the same images. Taking the Euclidean distance function to measure the distance between text features, we can define α-multilingual individual fairness by:

Definition 1 (Multilingual Individual Fairness)**.** Given a set of image-text pairs $\{(I, T)\}$, a multimodal model M satisfies α-multilingual individual fairness if for all (I, T), for languages L and L':

$$|S(I, T^{(L)}) - S(I, T^{(L')})| \leq \alpha \|\boldsymbol{t}^{(L)} - \boldsymbol{t}^{(L')}\|$$

where $\boldsymbol{t}^{(L)}$ is the textual representation vector yielded by M in language L.

Here, α is a parameter to control the ratio of similarity gap to the text feature vectors' distance, and smaller α indicates the model is individually fairer. Note that the similarity gap is at most 2, because the range of cosine similarity is $[0, 1]$. In general settings, $S(I, T)$ is measured by the cosine similarity between the encoded visual vector $\boldsymbol{v}$ and textual vector $\boldsymbol{t}$.

Lemma 1. *Denote* $\mathcal{O}_\rho(\boldsymbol{t}) = \{\boldsymbol{x} \mid \|\boldsymbol{x} - \boldsymbol{t}\| \leq \rho\}$ *to be a closed ball of radius* $\rho > 0$ *and center* $\boldsymbol{t}$*. Then* *for any visual representation vector $\boldsymbol{v}$,*

$$\sup_{\substack{\boldsymbol{t}^{(L')} \in \mathcal{O}_\rho(\boldsymbol{t}^{(L)}) \\ 0 \le \rho < \|\boldsymbol{t}^{(L)}\|}} |S(\boldsymbol{v}, \boldsymbol{t}^{(L')}) - S(\boldsymbol{v}, \boldsymbol{t}^{(L)})| \le \sqrt{2(1 - \sqrt{1 - (\frac{\rho}{\|\boldsymbol{t}^{(L)}\|})^2})} \quad (1)$$

where $S(\cdot, \cdot)$ denotes the cosine similarity, $\boldsymbol{t}^{(L)}$ and $\boldsymbol{t}^{(L')}$ are textual representation vectors for languages L and L', respectively.

We defer the proof to Appendix A.1. Lemma 1 implies that when the distance between multilingual textual representation vectors is bounded, the similarity with images can be bounded in terms of their distance. It is worth noting that the bounds are independent of the visual representation vectors. Nevertheless, the form of upper bound in Theorem 1 is a bit sophisticated, and can be simplified when $\rho \ll \|\boldsymbol{t}^{(L)}\|$.

Theorem 2. *When $\|\boldsymbol{t}^{(L')} - \boldsymbol{t}^{(L)}\| \ll \|\boldsymbol{t}^{(L)}\|$,*

$$|S(\boldsymbol{v}, \boldsymbol{t}^{(L')}) - S(\boldsymbol{v}, \boldsymbol{t}^{(L)})| \lessapprox \frac{\|\boldsymbol{t}^{(L')} - \boldsymbol{t}^{(L)}\|}{\|\boldsymbol{t}^{(L)}\|}.$$

Theorem 2 is a direct application of Lemma 1 when the distance between multilingual vectors is small enough, and extends in many natural cases to approximate the multilingual individual fairness with $\alpha \approx \frac{1}{\|\boldsymbol{t}^{(L)}\|}$. The proof can be found in Appendix A.2. Theorem 2 implicates to what degree the multimodal model satisfies individual fairness when text snippets are well aligned between different languages.

4 Multilingual Group Fairness

Distinct from individual fairness, multilingual group fairness appeals to the idea that multimodal models should achieve equivalent predictive performance across different languages. From the perspective of representations, it is hard to carry out this demand without well-defined tasks and metrics. Hence it is natural to ask how to define group fairness in this scenario properly? In this section, we shall answer this question by equating language as a unique dimension of group membership relating to the text modality. We formulate the criteria by equalizing the accuracy rates over different languages. We also observe that images are often connected to people in protected or unprotected groups. Given the image-text pairs, we consider the accuracy disparity across different languages conditioning the subgroup of images.

4.1 Equality of Accuracy across Languages

Given a dataset $\mathcal{D}$ consisting of ground-truth image-text pairs $\{(I_i, T_i)\}$ and each text can be in different languages. The goal of a multimodal model M is to predict the similarity $S(I_i, T_j)$ for any image I_i and text T_j. Then the model matches $\hat{T}_i$ for images I_i by selecting the text with highest similarity scores, i.e., $\hat{T}_i = \arg\max_j S(I_i, T_j)$.

$$\texttt{Acc}(M) = \frac{1}{|\mathcal{D}|} \sum_{\mathcal{D}} \mathbb{1}[\hat{T}_i = T_i] \quad (2)$$

We use the superscript (L) to indicate the accuracy $\texttt{Acc}^{(L)}$ is evaluated in language L. Next, we take language as group membership and define multilingual accuracy parity by equalizing accuracy across languages.

Definition 2 (multilingual accuracy parity)**.** A multimodal model M satisfies *multilingual accuracy parity* if $\texttt{Acc}^{(L)}(M) = \texttt{Acc}^{(L')}(M)$ for all languages L, L'.

In practice, it is impossible to achieve accuracy parity for all languages. Following (Hu et al., 2020), we use

$$\texttt{Gap}_M(L, L') = |\texttt{Acc}^{(L)}(M) - \texttt{Acc}^{(L')}(M)| \quad (3)$$

to represent the cross-lingual gap for model M.

4.2 When Language Meets Groups in Images

The above discussion on group fairness considers language as the sole group membership. In the real-world image and text applications, the people portrayed in the images are often associated with protected groups. For instance, the face attribute dataset (Liu et al., 2015) contains sensitive attributes, such as race, age and gender. Let G denote the group membership of images and $\mathcal{D}_a$ denote the subset of data examples $\mathcal{D}$ given $G = a$. The accuracy of a multimodal model evaluated on the images of subgroup a is defined as

$$\texttt{Acc}_a(M) = \frac{1}{|\mathcal{D}_a|} \sum_{\mathcal{D}_a} \mathbb{1}[\hat{T}_i = T_i] \quad (4)$$

When language is connected to images of different groups, we can define accuracy disparity between group a and group b with respect to model M within language L as

$$\texttt{Disp}_M^{(L)}(a, b) = |\texttt{Acc}_a^{(L)}(M) - \texttt{Acc}_b^{(L)}(M)| \quad (5)$$

$\texttt{Disp}$ represents the group rate gap in a single language. Mirroring *multilingual accuracy parity*, we can define the *multilingual group rate parity* as below.

Definition 3 (multilingual group rate parity)**.** A multi-modal model M satisfies multilingual group rate parity if $\texttt{Disp}_M^{(L)}(a,b) = \texttt{Disp}_M^{(L')}(a,b)$ with respect to groups a, b associated with images for all languages.

Definition 2 and Definition 3 evaluate the fairness of multilingual representations from diverse aspects. More broadly, we may be interested in the accuracy gap between different combinations of languages and groups. A common case is that there are only two protected groups (e.g. female and male, young and old). Let $p_a = \frac{|\mathcal{D}_a|}{|\mathcal{D}|}$ and $p_b = \frac{|\mathcal{D}_b|}{|\mathcal{D}|}$ represent the population proportions of group a and group b respectively, satisfying $p_a + p_b = 1$. Then we can decompose the cross-lingual cross-group accuracy disparity as below:

Proposition 3. *When there are only two protected groups a and b, the following inequality holds for any two languages L and L'*

$$|\texttt{Acc}_a^{(L)} - \texttt{Acc}_b^{(L')}| \leq \texttt{Gap}(L, L') \\ + p_b \cdot \texttt{Disp}^{(L)}(a,b) + p_a \cdot \texttt{Disp}^{(L')}(a,b) \quad (6)$$

The proof can be found in Appendix A.3. Proposition 3 guarantees that the accuracy disparity between any combinations of languages and protected groups can be upper bounded by a variety of factors, and implicates that we only need to focus on cross-lingual gap and group rate gap measures to assess multilingual group fairness. In what follows, we will take a closer look at how the multilingual CLIP model performs with compositions of languages and protected groups under these fairness criteria.

5 Evaluations

In this section, we work with the pre-trained multilingual CLIP (Carlsson and Ekgren, 2021) model to study multilingual fairness. We validate the extent to which the model is individually fair across different languages in Section 5.1. We characterize the prevalence of multilingual group unfairness on human faces in Section 5.2 and Section 5.3. These empirical evaluations shed light on potential directions for mitigating unfairness in multilingual multimodal representations.

5.1 Multilingual Individual Fairness

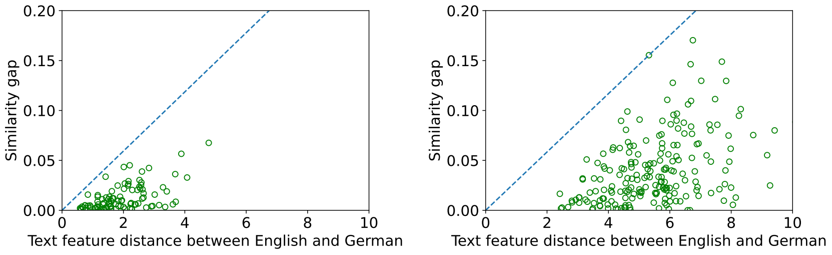
The theoretical analysis on multilingual individual fairness posed in Section 3 implies that the ratio of similarity difference to their text feature distance can be bounded by the reciprocal of the length of text feature vectors. To verify the implication, we conduct experiments on the Multi30K dataset (Elliott et al., 2016).

Dataset. The Multi30K dataset (Elliott et al., 2016) contains 31,014 Flickr30K (Young et al., 2014) images and composes the *translation* and the *independent* portions of English-German caption pairs. The German translations were collected from professional English-German translators by translating the English captions without seeing the images, one per image. The independent portion was independently annotated by German crowdworkers after seeing the images instead of English captions, five per image. Hence, the translated captions are strongly aligned in both languages, while the independent descriptions may have distinct context. We use 1,000 test images for our evaluation. For the independent portion, we select the first English caption and the first German caption of the five to pair with the image for a fair comparison.

Results. We embed each English-German caption pair into textual representation vectors and the corresponding image into visual representation vectors. We compute the Euclidean distance between English-German text features, as well as the cosine similarity with respect to the image features. We plot their cross-lingual gap on the translation and the independent portions in Fig. 1a and Fig. 1b, respectively. For both portions, the blue dashed lines represent the empirical upper bounds of the ratio between similarity gap and text feature distance.

Unsurprisingly, we find out that the English-German captions are more closely aligned on the translation portion (the average textual feature distance is 1.86) than the independent portion (average distance is 5.69). The similarity gaps regarding the translation portion are below 0.06 in general, and those regarding the independent portion are above 0.10 for many instances. The reason is apparent: translated captions have more similar semantics owning to the professional text-to-text translations, while independent captions have more diverse expressions of the same images, even if they might refer to the same content.

On the other hand, we observe that the slopes



(a) Translation

(b) Independent

	translation	independent
en	50.4	59.2
de	45.6	36.3
Gap	4.8	22.9

(c) Accuracy disparity

Figure 1: **We empirically examine how does the multilingual CLIP fare on the translation and the independent portions.** Fig. (a) and (b): the x-axis represents the distance between English and German captions, the y-axis represents the gap between their corresponding similarity scores, and the slope of blue dashed lines represents the empirical α for multilingual individual fairness. Fig. (c): we evaluate the accuracy for image-text matching, and find out that the independent portion incurs huge accuracy disparity compared with the translation portion.

of blue dashed lines in Fig. 1a and Fig. 1b are approximate to each other, i.e., the empirical α for both portions are similar. This fact implies that the multilingual CLIP model evaluated on two different text corpora share a similar level of individual fairness, even though the cross-lingual similarity gaps are quite different. We also note that the empirical upper bound of α are much smaller than the theoretical upper bound $\frac{1}{\|\boldsymbol{t}^{(L)}\|}$ in Theorem 2.

Although we have verified that multilingual multimodal representations satisfy similar individual fairness, we demonstrate that they violate group fairness by evaluating their image-text matching accuracy. As shown in Fig. 1c, English captions dominate the Top-1 image-text matching accuracy over German captions, with 4.8% higher on the translation portion and 22.9% higher on the independent portion. This observation delivers an important message for researchers who are interested in learning fair representations (Ruoss et al., 2020): individual fairness *does not* flatly prevent accuracy disparity among different languages (Binns, 2020).

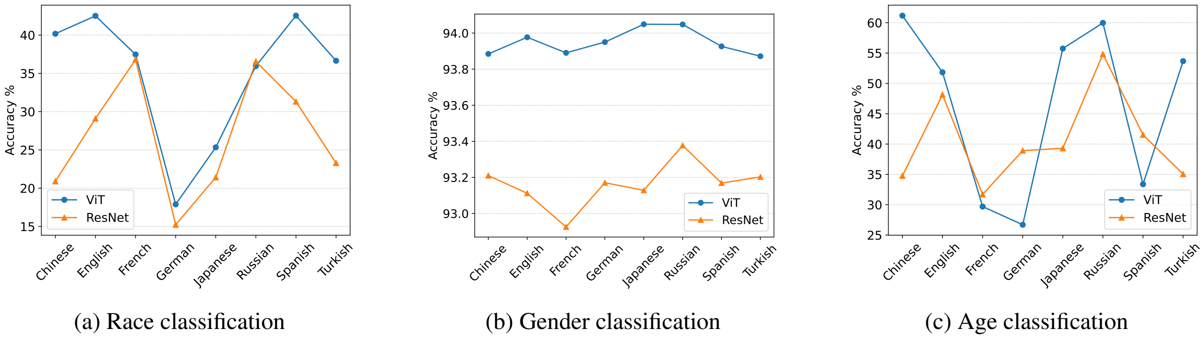
5.2 Multilingual Accuracy Disparity

Dataset. FairFace (Karkkainen and Joo, 2021) is a face attribute dataset for the balanced race, gender, and age groups. It categorizes gender into two groups, including female and male, and race into seven groups, including White, Black, Indian, East Asian, Southeast Asian, Middle Eastern, and Latino. For ages, we categorize the raw labels into five groups: infants (0–2), children and adolescents (2–19), adults (20–49), middle age adults (50–69), and seniors (more than 70). We follow their original data split and select the validation set consisting of 10,940 face images for evaluation.

Languages. We analyze the multilingual group fairness for 8 languages: Chinese (zh), English (en), French (fr), German (de), Japanese (ja), Russian (ru), Spanish (es), and Turkish (tr). We select English as the pivot language and write natural language prompts in English. Then we translate them into other languages: we first use Google Translate and then recruit native speakers to rate the prompts and fix any potential errors on Amazon Mechanical Turk (see Appendix D for more details). The rationale for only using English as the pivot language is that the multilingual CLIP (Carlsson and Ekgren, 2021) selects English as the pivot language for aligning multilingual text embeddings.

Text Prompts. Following Radford et al. (2021), we construct the text prompt by the template "A photo of a {label} person". Concretely, for gender classification, we construct the text prompt "A photo of a woman" when the gender attribute is female, and construct "A photo of a man" otherwise. For race classification, we construct the text prompt by "A photo of a(n) {race} person". Note that Indian actually refers to South Asian ethnic groups in the Fairface race taxonomy (Karkkainen and Joo, 2021) but it can refer to Native Americans as well. To avoid ambiguity, we replace "Indian" by "South Eastern" to construct the prompts. For age classification, we notice that the age attributes in Fairface dataset are numeric values and use the template "A photo of a person aged {age} years" to construct text prompts.

Results. We probe the multilingual accuracy disparity for race classification, gender classification, and age classification, as shown in Fig. 2. We use two different pre-trained image encoders for extracting visual representation vectors, including



(a) Race classification

(b) Gender classification

(c) Age classification

Figure 2: **Race, gender, and age classification accuracy across different languages.** The languages are in alphabetical order. Two different vision encoders for encoding image features are evaluated, including Vision Transformer (ViT) (Dosovitskiy et al., 2021) and ResNet-50 (ResNet) (He et al., 2016).

Vision Transformer (Dosovitskiy et al., 2021) and ResNet-50 (He et al., 2016). We observe that:

- **Cross-lingual gap varies across different protected groups.** The predictive accuracy for gender classification is consistently higher than 90% across all the languages. In contrast, the multimodal model has relatively poor performance and more considerable variance for race and age classification. Furthermore, race classification yields 24.66% accuracy disparity and age classification yields 34.47% accuracy disparity for Vision Transformer. This implies that the huge disparity may result from the poor predictive performance of the model.
- **Visual representations affect accuracy disparity.** For race classification, Vision Transformer features generally achieve higher accuracy across all languages than ResNet-50 (34.82% *vs.* 26.83% on average) except for Russian. The standard deviation of Vision Transformer is higher than ResNet-50 (8.18% *vs.* 7.34%). The maximal accuracy gap for Vision Transformer is 30.40% between German and Spanish, while the maximal accuracy gap for ResNet-50 is 23.12% between German and French. For gender classification, Vision Transformer dominantly achieves higher accuracy and incurs less accuracy gap. For age classification, the accuracy is moderately low for all languages. However, Vision Transformer has 63.1% accuracy in Chinese while only 25.8% accuracy in German, exaggerating the accuracy gap between languages.

In Table 1, we present the complete results of Fig. 3 by compositions of gender and race groups across different languages.

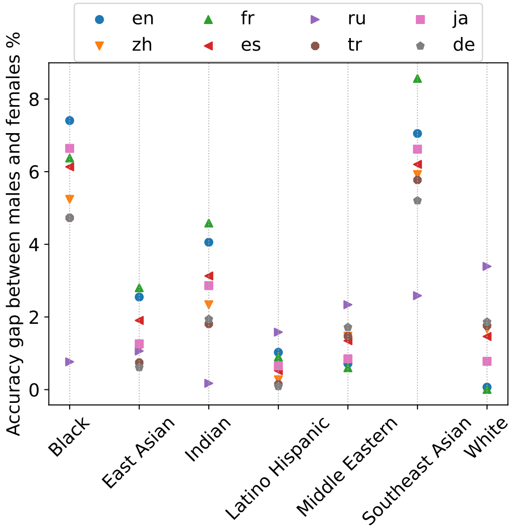


Figure 3: **Gender accuracy gap across different languages and racial groups.** Black and Southeast Asian people face significant larger gender gaps than other racial groups in most languages.

5.3 Multilingual Group Rate Disparity

We evaluate multilingual group rate disparity for gender classification on Fairface dataset. We follow the same setup as described in Section 5.2 and measure the gender gap given by Eq. (5), where a is the composition of male and various race groups, b is the composition of female and various race groups. We defer the complete results to Table 1 in Appendix B. We try to answer the following research questions:

- **How do gender gaps differ across protected groups?** We plot the gender accuracy gap across different languages and racial groups in Fig. 3. It is clearly shown that Black and Southeast Asian groups dominantly exhibit larger gender gaps than other groups. We also observe that French has a similar performance with English. We con-

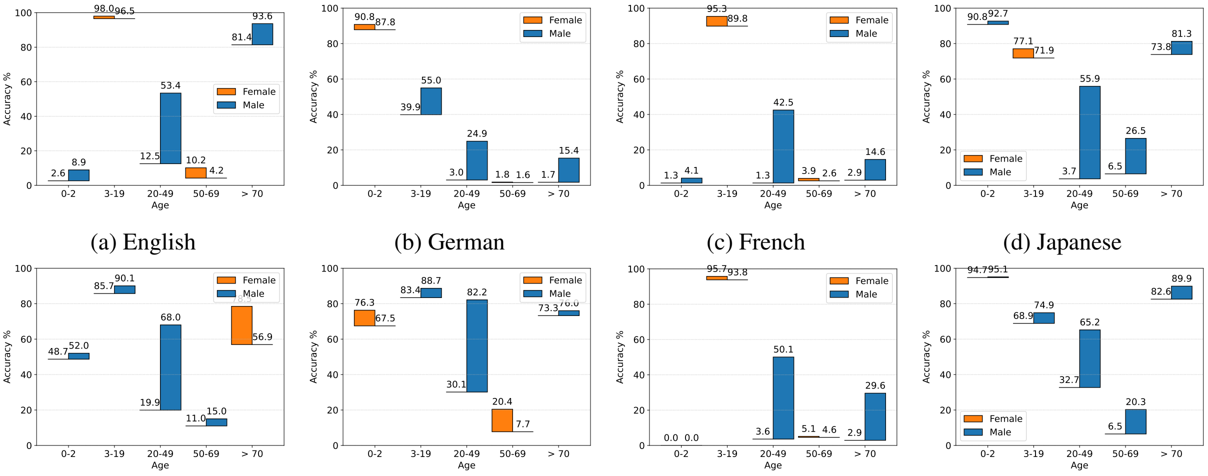
Table 1: **Gender classification accuracy of FairFace images by race groups across different languages.** We note the maximal gender gap across races with underline and the maximal gender gap across languages in **bold**. Taking English as the pivot language, we also highlight any amplified gap compared to English in red and any mitigated gap in green.

					East	Southeast	Middle		
Language	Gender	White	Black	Indian	Asian	Asian	Eastern	Latino	Average
English	Female	95.1	90.9	94.5	95.2	96.0	96.0	94.2	94.6
	Male	95.2	83.5	90.4	92.7	89.0	96.7	93.2	91.5
	`Disp`	0.1	**7.4**	4.1	2.5	7.0	0.7	1.0	3.0
German	Female	93.8	90.1	94.0	94.2	95.0	95.5	93.9	93.8
	Male	95.6	85.4	92.0	93.6	89.8	97.2	93.9	92.5
	`Disp`	1.9	4.7	1.9	0.6	5.2	1.7	0.1	1.3
French	Female	95.0	90.4	94.6	95.0	96.3	95.7	94.2	94.5
	Male	95.0	84.0	90.0	92.1	87.8	96.3	93.3	91.2
	`Disp`	0.0	6.4	**4.6**	**2.8**	**8.6**	0.6	0.9	**3.2**
Japanese	Female	94.5	90.6	94.4	94.7	95.7	95.7	94.1	94.2
	Male	95.3	84.0	91.5	93.4	89.1	96.6	93.4	91.9
	`Disp`	0.8	6.6	2.9	1.3	6.6	0.8	0.7	2.3
Turkish	Female	93.9	90.0	93.8	94.6	95.3	95.5	94.1	93.9
	Male	95.6	85.2	92.0	93.8	89.5	96.9	93.9	92.4
	`Disp`	1.8	4.7	1.8	0.7	5.8	1.5	0.1	1.4
Russian	Female	93.0	88.4	93.1	93.4	94.6	95.2	93.4	93.0
	Male	96.4	87.6	93.2	94.5	92.0	97.5	95.0	93.7
	`Disp`	**3.4**	0.8	0.2	1.1	2.6	**2.3**	**1.6**	0.7
Spainish	Female	94.1	90.5	94.4	95.1	95.6	95.5	94.2	94.2
	Male	95.5	84.4	91.2	93.2	89.4	96.8	93.7	92.0
	`Disp`	1.5	6.1	3.1	1.9	6.2	1.3	0.5	2.2
Chinese	Female	93.9	90.1	94.1	94.8	95.4	95.5	94.2	94.0
	Male	95.5	84.9	91.8	93.7	89.5	96.9	93.9	92.3
	`Disp`	1.7	5.2	2.3	1.1	5.9	1.5	0.3	1.7

jecture this is because English and French share the same alphabet and similar syntactic structures. Besides, as shown in Table 1, English and French have the largest race inequality regarding gender gap—nearly zero gender gaps for White but near the maximal gaps for Black.

- **Are gender gaps amplified for different languages when compared with English?** We report the accuracy gap on gender classification of FairFace images by race groups across different languages in Table 1. We take English as the pivot language and examine whether the accuracy gaps by race groups are amplified for other languages. Compared with English, accuracy gaps for White and Middle Eastern groups are generally amplified for other languages. On the other hand, accuracy gaps are generally mitigated for groups including Black, Indian, East Asian, Southeast Asian, and Latino groups. The averaged cross-lingual gaps are mitigated for all the languages except for French.

We also evaluate multilingual group rate disparity for age classification. We composite gender and age as the group membership. We plot the age classification accuracy by female and male groups across different languages in Fig. 4. The blue bars indicate that the male group has higher accuracy than the female group, while the orange bars indicate that the female group has higher accuracy than the male group. The heights of bars represent the accuracy gaps between male and female groups. In general, the male group has higher accuracy than the female group. Especially, adults (20–49 years old) consistently suffer huge gender gaps across all the languages, with the largest gap 52.2% for Japanese. It is worth noting that the numerals to express ages are identical in text prompts for different languages, e.g., “a person aged 20 to 49 years” in English versus “eine Person im Alter von 20 bis 49 Jahren” in German. This controlled experiment



(a) English (b) German (c) French (d) Japanese

(e) Turkish (f) Russian (g) Spanish (h) Chinese

Figure 4: **Age classification accuracy across female and male groups for different languages.** The blue bars indicate that the male group has higher accuracy than the female group, while orange bars indicate that the female group has higher accuracy. The heights of bars represent the accuracy gaps between male and female groups.

helps us better understand whether the identical numeric digits have distinct meanings in multilingual contexts. As shown in Fig. 4, although text prompts in different languages share the same numerals of ages, the yielding accuracy exhibits significant disparity across languages. One prominent example is that the predictive accuracy for infants (0–2 years old) is 5.8% for English and 2.6% for French, but 89.4% for German and 91.6% for Japanese, implying the presence of significant cross-lingual accuracy gaps.

6 Conclusion and Limitation

Our work extends a growing body of fairness discourse in multilingual and multimodal learning to explore how the multilingual fairness notions, characterized by individual fairness and group fairness, are formulated on the multimodal representations. We stress that multimodal representations are individually fair, but do not prevent accuracy disparity across groups. Our extensive experimental results reveal the negative impacts caused by carelessly applying pre-trained general-purpose multimodal representations. Just one example of this, as discussed in Section 5.3, is the significant disparities between cross-lingual gender gaps occurred in age classification. We believe the findings and insights gained through this work can encourage future work to investigate how to mitigate multidimensional biases in representation learning and prevent disparities in the downstream decision-making process.

Our work also has limitations. This work does not provide a thorough explanation on whether the biases and disparities result from the multilingual model itself, or from the datasets it is pre-trained on. However, to give a convincing explanation, it requires either access to large amounts of training data with privacy concerns (the complete datasets for training CLIP are not released yet), or ample computational resources for reproducing the training process. This research question itself is important and worth investigating further.

Broader Impact

This work provides insights into fairness in the context of multilingual and multimodal representations. We recognize potential ethical concerns that may arise in the evaluation and address them below.

Firstly, the empirical evaluation for multilingual group fairness adopts the categories of protected groups introduced in the FairFace dataset (Karkkainen and Joo, 2021). We are aware that gender can be non-binary, and individuals can be self-identified outside male and female. Some terms of race attributes in the dataset, such as Latino and Hispanic, are rooted in culture and ethnicity and should not be treated as racial categories. In addition, facial images of low population groups, including Hawaiian and Pacific Islanders and Native Americans, are discarded during data collection. The sensitive attributes in the original FairFace datasets are identified and annotated by human crowd workers. It is possible that the labels of gender, race, and age contain implicit biases and noises. However, these ethical concerns arise from the data collection of the FairFace dataset per se. We anticipate that the methodology we adopted to study multilingual fairness can still be generalized to other data source when more inclusive data collections are available.

Secondly, image classification on the FairFace dataset relies on human-crafted text prompts. The fashion of prompt engineering can be dated from pre-training image and text representations with natural language supervision (Li et al., 2017; Radford et al., 2021). To avoid offensive and harmful speech towards certain protected groups, we construct the text prompts in a descriptive intent and a neutral tone.

Finally, the intention for performing classification with sensitive attributes is to validate the presence of biases in pre-trained representations rather than to acquire the personal information of people in the images. Both the evaluated pre-trained models and the benchmark datasets are publicly accessible, and we carefully follow their licenses and agreements for usage. In this sense, we do not foresee any data privacy or information security issues.

Acknowledgements

The authors would like to thank Jieyu Zhao, Kai-Wei Chang, Lise Getoor, Jess Law, Hao Tan, and the anonymous reviewers for their fruitful discussions and constructive feedback. This work is supported by the UCSC Applied Artificial Intelligence Initiative (AAII) and the National Science Foundation (NSF) under grant IIS-2040800.

References

Reuben Binns. 2020. On the apparent conflict between individual and group fairness. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, FAT* '20, page 514–524, New York, NY, USA. Association for Computing Machinery.

Andrea Burns, Donghyun Kim, D. Wijaya, Kate Saenko, and Bryan A. Plummer. 2020. Learning to scale multilingual representations for vision-language tasks. *ArXiv*, abs/2004.04312.

Kaylee Burns, Lisa Anne Hendricks, Trevor Darrell, and Anna Rohrbach. 2018. Women also snowboard: Overcoming bias in captioning models. In *ECCV*.

Fredrik Carlsson and Ariel Ekgren. 2021. Pre-trained multilingual-clip encoders. `https://github.com/FreddeFrallan/Multilingual-CLIP`.

Yen-Chun Chen, Linjie Li, Licheng Yu, Ahmed El Kholy, Faisal Ahmed, Zhe Gan, Yu Cheng, and Jingjing Liu. 2020. Uniter: Universal image-text representation learning. In *ECCV*.

Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. 2021. Unifying vision-and-language tasks via text generation. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 1931–1942. PMLR.

Monojit Choudhury and Amit Deshpande. 2021. How linguistically fair are multilingual pre-trained language models? In *AAAI-21*. AAAI, AAAI.

A. Chouldechova. 2017. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5 2:153–163.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *ACL*.

Karan Desai and Justin Johnson. 2021. VirTex: Learning Visual Representations from Textual Annotations. In *CVPR*.

J. Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT*.

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 *International Conference on Learning Representations*.

Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. 2012. Fairness through awareness. In *Proceedings of the 3rd Innovations in Theoretical Computer Science Conference*, ITCS '12, page 214–226, New York, NY, USA. Association for Computing Machinery.

Desmond Elliott, Stella Frank, Khalil Sima'an, and Lucia Specia. 2016. Multi30K: Multilingual English-German image descriptions. In *Proceedings of the 5th Workshop on Vision and Language*, pages 70–74, Berlin, Germany. Association for Computational Linguistics.

Zhe Gan, Yen-Chun Chen, Linjie Li, Chen Zhu, Yu Cheng, and Jingjing Liu. 2020. Large-scale adversarial training for vision-and-language representation learning. In *NeurIPS*.

Ana Valeria González, Maria Barrett, Rasmus Hvingelby, Kellie Webster, and Anders Søgaard. 2020. Type B reflexivization as an unambiguous testbed for multilingual multi-task gender bias. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2637–2648, Online. Association for Computational Linguistics.

Moritz Hardt, Eric Price, and Nathan Srebro. 2016. Equality of opportunity in supervised learning. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, NIPS'16, page 3323–3331, Red Hook, NY, USA. Curran Associates Inc.

Kaiming He, X. Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. *2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 770–778.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. XTREME: A massively multilingual multitask benchmark for evaluating cross-lingual generalisation. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 4411–4421. PMLR.

Xiaolei Huang, Linzi Xing, Franck Dernoncourt, and Michael J. Paul. 2020. Multilingual Twitter corpus and baselines for evaluating demographic bias in hate speech recognition. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 1440–1448, Marseille, France. European Language Resources Association.

Kimmo Karkkainen and Jungseock Joo. 2021. Fairface: Face attribute dataset for balanced race, gender, and age for bias measurement and mitigation. In *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*, pages 1548–1558.

Iro Laina, C. Rupprecht, and N. Navab. 2019. Towards unsupervised image captioning with shared multimodal embeddings. *2019 IEEE/CVF International Conference on Computer Vision (ICCV)*, pages 7413–7423.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. In *NeurIPS*.

Ang Li, A. Jabri, Armand Joulin, and Laurens van der Maaten. 2017. Learning visual n-grams from web data. *2017 IEEE International Conference on Computer Vision (ICCV)*, pages 4193–4202.

Gen Li, Nan Duan, Yuejian Fang, Daxin Jiang, and M. Zhou. 2020a. Unicoder-vl: A universal encoder for vision and language by cross-modal pre-training. In *AAAI*.

Xiujun Li, Xi Yin, C. Li, X. Hu, Pengchuan Zhang, Lei Zhang, Longguang Wang, H. Hu, Li Dong, Furu Wei, Yejin Choi, and Jianfeng Gao. 2020b. Oscar: Object-semantics aligned pre-training for vision-language tasks. In *ECCV*.

Ziwei Liu, Ping Luo, Xiaogang Wang, and Xiaoou Tang. 2015. Deep learning face attributes in the wild. In *Proceedings of International Conference on Computer Vision (ICCV)*.

Jiasen Lu, Dhruv Batra, Devi Parikh, and Stefan Lee. 2019. Vilbert: Pretraining task-agnostic visiolinguistic representations for vision-and-language tasks. In *NeurIPS*.

Aditya Mogadala, Marimuthu Kalimuthu, and Dietrich Klakow. 2021. Trends in integration of vision and language research: A survey of tasks, datasets, and methods. *J. Artif. Int. Res.*, 71:1183–1317.

Arvind Narayanan. 2018. Translation tutorial: 21 fairness definitions and their politics. In *Proc. Conf. Fairness Accountability Transp.*

Telmo Pires, Eva Schlinger, and Dan Garrette. 2019. How multilingual is multilingual BERT? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4996–5001, Florence, Italy. Association for Computational Linguistics.

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*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR.

Aditya Ramesh, Mikhail Pavlov, Gabriel Goh, Scott Gray, Chelsea Voss, Alec Radford, Mark Chen, and Ilya Sutskever. 2021. Zero-shot text-to-image generation. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 8821–8831. PMLR.

Candace Ross, B. Katz, and Andrei Barbu. 2021. Measuring social biases in grounded vision and language embeddings. In *NAACL*.

Anian Ruoss, Mislav Balunovic, Marc Fischer, and Martin Vechev. 2020. Learning certified individually fair representations. In *Advances in Neural Information Processing Systems 33*.

Tejas Srinivasan and Yonatan Bisk. 2021. Worst of both worlds: Biases compound in pre-trained vision-and-language models. *ArXiv*, abs/2104.08666.

Weijie Su, Xizhou Zhu, Yue Cao, Bin Li, Lewei Lu, Furu Wei, and Jifeng Dai. 2020. Vl-bert: Pretraining of generic visual-linguistic representations. In *International Conference on Learning Representations*.

Hao Tan and Mohit Bansal. 2019. Lxmert: Learning cross-modality encoder representations from transformers. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*.

Ruixiang Tang, Mengnan Du, Yuening Li, Zirui Liu, Na Zou, and Xia Hu. 2021. Mitigating gender bias in captioning systems. In *Proceedings of the Web Conference 2021*, WWW '21, page 633–645, New York, NY, USA. Association for Computing Machinery.

Nam S. Vo, Lu Jiang, C. Sun, K. Murphy, L. Li, Li Fei-Fei, and James Hays. 2019. Composing text and image for image retrieval - an empirical odyssey. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 6432–6441.

Jialu Wang, Yang Liu, and Xin Eric Wang. 2021. Are gender-neutral queries really gender-neutral? mitigating gender bias in image search. In *EMNLP*.

Shijie Wu and Mark Dredze. 2020. Are all languages created equal in multilingual bert? In *REPL4NLP*.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Transactions of the Association for Computational Linguistics*, 2:67–78.

Fei Yu, Jiji Tang, Weichong Yin, Yu Sun, Hao Tian, Hua Wu, and Haifeng Wang. 2021. Ernie-vil: Knowledge enhanced vision-language representations through scene graphs. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 3208–3216. AAAI Press.

Muhammad Bilal Zafar, I. Valera, M. Gomez-Rodriguez, and K. Gummadi. 2017. Fairness constraints: Mechanisms for fair classification. In *AISTATS*.

R. Zemel, Ledell Yu Wu, Kevin Swersky, T. Pitassi, and C. Dwork. 2013. Learning fair representations. In *ICML*.

Jieyu Zhao, Subhabrata Mukherjee, Saghar Hosseini, Kai-Wei Chang, and Ahmed Hassan Awadallah. 2020. Gender bias in multilingual embeddings and cross-lingual transfer. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2896–2907, Online. Association for Computational Linguistics.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2017. Men also like shopping: Reducing gender bias amplification using corpus-level constraints. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2979–2989, Copenhagen, Denmark. Association for Computational Linguistics.

L. Zhou, H. Palangi, Lei Zhang, H. Hu, Jason J. Corso, and Jianfeng Gao. 2020. Unified vision-language pre-training for image captioning and vqa. In *AAAI*.

Zhaowei Zhu, Tianyi Luo, and Yang Liu. 2022. The rich get richer: Disparate impact of semi-supervised learning. In *International Conference on Learning Representations*.

A Omitted Proofs

A.1 Proof of Lemma 1

Proof. Given by the definition of cosine similarity, we have

$$\begin{aligned}&|S(\boldsymbol{v},\boldsymbol{t}^{(L)})-S(\boldsymbol{v},\boldsymbol{t}^{(L')})|\\ =&|\frac{\boldsymbol{v}\cdot\boldsymbol{t}^{(L)}}{\|\boldsymbol{v}\|\|\boldsymbol{t}^{(L)}\|}-\frac{\boldsymbol{v}\cdot\boldsymbol{t}^{(L')}}{\|\boldsymbol{v}\|\|\boldsymbol{t}^{(L')}\|}|\\ =&\frac{|\boldsymbol{v}\cdot(\|\boldsymbol{t}^{(L')}\|\boldsymbol{t}^{(L)}-\|\boldsymbol{t}^{(L)}\|\boldsymbol{t}^{(L')})|}{\|\boldsymbol{v}\|\|\boldsymbol{t}^{(L)}\|\|\boldsymbol{t}^{(L')}\|}\end{aligned}\tag{7}$$

From the definition of dot product,

$$\begin{aligned}&|\boldsymbol{v}\cdot(\|\boldsymbol{t}^{(L')}\|\boldsymbol{t}^{(L)}-\|\boldsymbol{t}^{(L)}\|\boldsymbol{t}^{(L')})|\le\\ &\qquad\|\boldsymbol{v}\|\cdot\|(\|\boldsymbol{t}^{(L')}\|\boldsymbol{t}^{(L)}-\|\boldsymbol{t}^{(L)}\|\boldsymbol{t}^{(L')})\|\end{aligned}\tag{8}$$

We plug Eq. (8) into Eq. (7) and eliminate the variable $\boldsymbol{v}$

$$\begin{aligned}&|S(\boldsymbol{v},\boldsymbol{t}^{(L)})-S(\boldsymbol{v},\boldsymbol{t}^{(L')})|\le\\ &\qquad\frac{\|(\|\boldsymbol{t}^{(L')}\|\boldsymbol{t}^{(L)}-\|\boldsymbol{t}^{(L)}\|\boldsymbol{t}^{(L')})\|}{\|\boldsymbol{t}^{(L)}\|\|\boldsymbol{t}^{(L')}\|}\end{aligned}\tag{9}$$

Let θ denote the angle between $\boldsymbol{t}^{(L)}$ and $\boldsymbol{t}^{(L')}$, i.e.,

$$\cos\theta=\frac{\boldsymbol{t}^{(L)}\cdot\boldsymbol{t}^{(L')}}{\|\boldsymbol{t}^{(L)}\|\|\boldsymbol{t}^{(L')}\|},$$

the square of numerator in Eq. (9) expands as

$$\begin{aligned}&(\|\boldsymbol{t}^{(L')}\|\boldsymbol{t}^{(L)}-\|\boldsymbol{t}^{(L)}\|\boldsymbol{t}^{(L')})^2\\ &\qquad=2\|\boldsymbol{t}^{(L)}\|^2\|\boldsymbol{t}^{(L')}\|^2(1-\cos\theta)\end{aligned}\tag{10}$$

Substituting the square root of Eq. (10) into Eq. (9), we eliminate the denominator and obtain

$$|S(\boldsymbol{v},\boldsymbol{t}^{(L)})-S(\boldsymbol{v},\boldsymbol{t}^{(L')})|\le\sqrt{2(1-\cos\theta)}\tag{11}$$

Recall that $\boldsymbol{t}^{(L')}\in\mathcal{O}_\rho(\boldsymbol{t}^{(L)})$, we can bound θ by the law of sines

$$\sup_\theta|\sin\theta|=\sup_{\boldsymbol{t}^{(L')}}\frac{\|\boldsymbol{t}^{(L')}-\boldsymbol{t}^{(L)}\|}{\|\boldsymbol{t}^{(L)}\|}=\frac{\rho}{\|\boldsymbol{t}^{(L)}\|}\tag{12}$$

Taking supremums on both sides of Eq. (11) and combining Eq. (12), we complete the proof

$$\begin{aligned}&\sup_{\substack{\boldsymbol{t}^{(L')}\in\mathcal{O}_\rho(\boldsymbol{t}^{(L)})\\0\le\rho<\|\boldsymbol{t}^{(L)}\|}}|S(\boldsymbol{v},\boldsymbol{t}^{(L')})-S(\boldsymbol{v},\boldsymbol{t}^{(L)})|\\ \le&\sup_\theta\sqrt{2(1-\sqrt{1-sin^2\theta})}\\ =&\sqrt{2(1-\sqrt{1-(\frac{\rho}{\|\boldsymbol{t}^{(L)}\|})^2})}\end{aligned}$$

□

A.2 Proof of Theorem 2

Proof. Due to Half-Angle Identities, Eq. (11) derives as

$$|S(\boldsymbol{v},\boldsymbol{t}^{(L')})-S(\boldsymbol{v},\boldsymbol{t}^{(L)})|\le2|\sin\frac{\theta}{2}|\tag{13}$$

For sufficiently small θ, i.e., $\|\boldsymbol{t}^{(L')}-\boldsymbol{t}^{(L)}\|\ll\|\boldsymbol{t}^{(L)}\|$, we take the first-order Taylor approximation

$$2|\sin\frac{\theta}{2}|\approx|\theta|\approx|\sin\theta|=\frac{\|\boldsymbol{t}^{(L')}-\boldsymbol{t}^{(L)}\|}{\|\boldsymbol{t}^{(L)}\|}\tag{14}$$

Combining Eq. (13) and Eq. (14) we complete the proof. □

A.3 Proof of Theorem 3

Proof. Expanding $|\mathtt{Acc}_a^{(L)}-\mathtt{Acc}_b^{(L')}|$ by triangle inequality we have

$$\begin{aligned}&|\mathtt{Acc}_a^{(L)}-\mathtt{Acc}_b^{(L')}|\\ =&|\mathtt{Acc}_a^{(L)}-\mathtt{Acc}^{(L)}+\mathtt{Acc}^{(L)}\\ &\quad-\mathtt{Acc}^{(L')}+\mathtt{Acc}^{(L')}-\mathtt{Acc}_b^{(L')}|\\ \le&|\mathtt{Acc}_a^{(L)}-\mathtt{Acc}^{(L)}|+|\mathtt{Acc}^{(L)}-\mathtt{Acc}^{(L')}|\\ &\quad+|\mathtt{Acc}^{(L')}-\mathtt{Acc}_b^{(L')}|\end{aligned}\tag{15}$$

Noticing that $\mathtt{Acc}^{(L)}=p_a\cdot\mathtt{Acc}_a^{(L)}+p_b\cdot\mathtt{Acc}_b^{(L)}$ and $p_a+p_b=1$, we have

$$\begin{aligned}&|\mathtt{Acc}_a^{(L)}-\mathtt{Acc}^{(L)}|\\ =&p_b\cdot|\mathtt{Acc}_a^{(L)}-\mathtt{Acc}_b^{(L)}|\\ =&p_b\cdot\mathtt{Disp}^{(L)}(a,b)\end{aligned}\tag{16}$$

Similarly,

$$\begin{aligned}&|\mathtt{Acc}^{(L')}-\mathtt{Acc}_b^{(L')}|\\ =&p_a\cdot|\mathtt{Acc}_a^{(L')}-\mathtt{Acc}_b^{(L')}|\\ =&p_a\cdot\mathtt{Disp}^{(L)}(a,b)\end{aligned}\tag{17}$$

Substituting Eq. (3), Eq. (16), and Eq. (17) into Eq. (15) we complete the proof. □

B Additional Experimental Results

B.1 Empirical Evaluation with Dissimilar Images and Text

We note that the theoretical analysis posed in Theorem 2 does *not* presume how the images are similar to the text. However, the evaluation in Section 5.1 only focuses on similar images and text. To complement for evaluation on dissimilar images and text, we measured to what extent the pre-trained model satisfies multilingual individual fairness for dissimilar images and captions in the Appendix B.1. Specifically, we randomly shuffle the images in the data set such that each image is paired with a random pair of English and German captions. Then we compare the similarity gaps between English and German captions with the images in terms of the encoded textual vector distance between English and German. We observe the same trends for dissimilar images and text: (1) The translation portion generally induces a smaller similarity gap than the independent portion. (2) The CLIP model evaluated on both text corpora has similar empirical α values.

C Computation Infrastructure

We use a GPU server with 4 NVIDIA RTX 2080 Ti GPUs for evaluation.

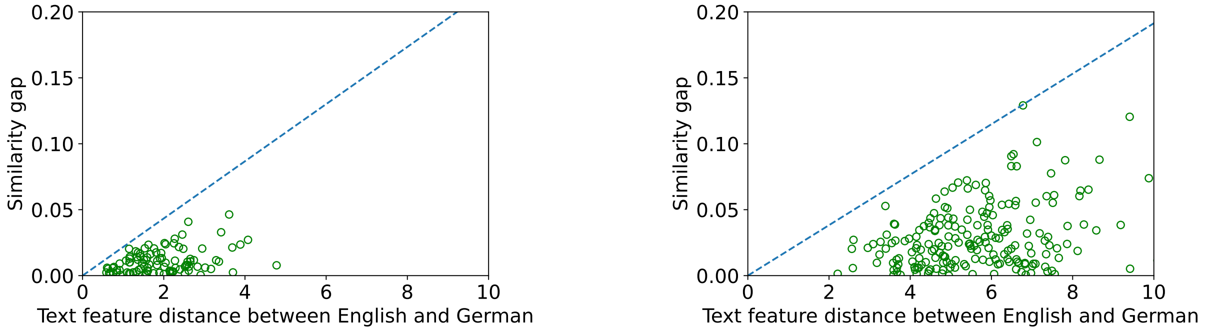
D Human Evaluation of the Quality of Machine Translated Text Promts

We recruited crowd workers at Amazon Mechanical Turk (AMT)[1] to evaluated the quality of text prompts generated in Section 5.2. The crowd workers were supposed to speak both the original language and the translated language to be qualified for completing the tasks. Each task contained one pair of text prompts in the original language (English) and the translated language and was assigned to at least five crowd workers. Each crowd worker was asked to rate the quality of translation from adequacy and fluency on a scale of 1–5. Specifically, we asked the crowd workers the following questions:

- **Adequacy:** does the translated text adequately expresses the meaning in the original text in English?
- **Fluency**: how good the translated language is?

We also asked the workers to point out and fix any potential problems in the prompts. We collected and visualized the crowdsourced ratings in Fig. 6. For Chinese, French, German, and Japanese, the crowd workers considered the translated text can adequately express all the meanings retained in the English prompts and is flawless. For Japanese, Russian, and Turkish, the crowd workers considered the translations can convey most of the message in the English prompts and are good in fluency.

[1] https://www.mturk.com/



(a) Translation

(b) Independent

Figure 5: **We empirically examine how does the multilingual CLIP fare on the translation and the independent portions.** Fig. (a) and (b): the x-axis represents the distance between English and German captions, the y-axis represents the gap between their corresponding *dissimilarity* scores, and the slope of blue dashed lines represents the empirical α for multilingual individual fairness.

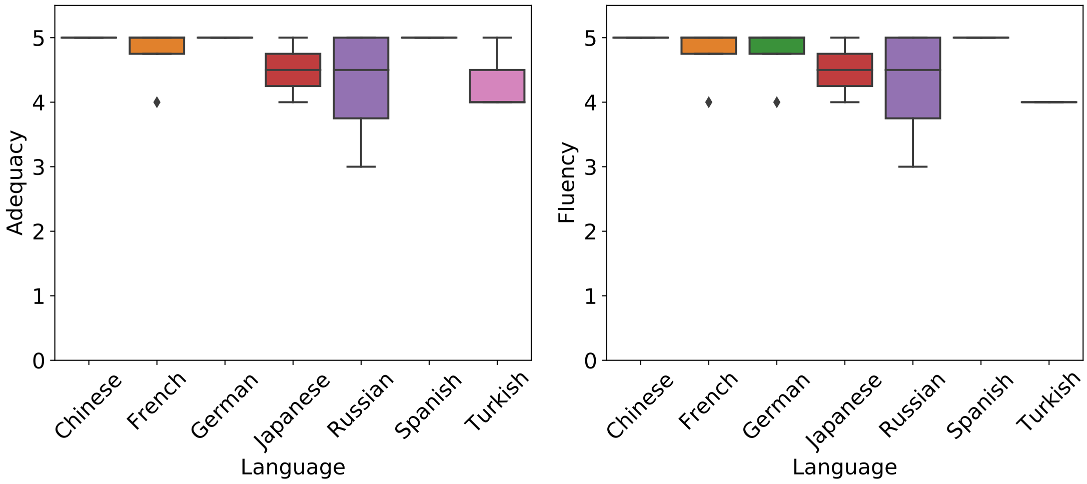


Figure 6: We recruited crowd workers at AMT to rate the adequacy and fluency of the machine translated text prompts on a scale of 1–5.