Università degli Studi di Milano-Bicocca
Dipartimento di Informatica, Sistemistica e Comunicazione
Corso di Laurea Magistrale in Data Science

Exposing Bias in Vision-Language Models

Relatore: Prof. Elisabetta Fersini
Correlatore: Prof. Albert Gatt

Relazione della prova finale di:
Fabrizio Cominetti
Matricola 882737

Anno Accademico 2022-2023

Contents

1 Introduction

In recent years, the integration of vision and language information has driven unprecedented advancements in data science, machine learning and artificial intelligence, revolutionizing fields ranging from computer vision to natural language processing. Vision and Language models are capable of processing, understanding and managing visual and textual data, and they have demonstrated remarkable proficiency in various tasks related to these modalities, such as image captioning, visual question answering, multimodal retrieval, image-text matching, and much more. However, among these impressive achievements, there is a significant concern: the possible existence and presence of bias within these models. Bias, whether conscious or unconscious, can manifest in various forms, with particular consideration given to racial and gender stereotypes alongside cultural prejudices. In the context of vision and language models, bias can significantly impact the fairness, reliability, and inclusivity of their outputs, potentially perpetuating societal inequalities and reinforcing harmful stereotypes. For these reasons it is crucial to comprehend and address bias in these models to guarantee fairness and ethics in artificial intelligence systems.

This master's thesis aims to investigate the intricate relationship between bias and vision-language models, exploring its manifestations, causes, and implications. Through a comprehensive analysis of three well-known pre-trained models, namely ALBEF, BLIP-2, and CLIP, the objective is to explore the status of bias and its effects on vision-language models.

The investigation begins with the construction of a novel dataset - a morphed dataset derived from the publicly available UTKFace dataset - which serves as a foundation for investigating gender and racial biases within vision and language processing. Then, a series of experiments will be conducted to explore how these biases affect different vision-language models. From zero-shot retrieval to zero-shot classification tasks, the goal is to uncover, identify and measure any potential bias, related to our categories of interest, in these experiments. Throughout the investigation, some fairness metrics will

be employed to quantitatively assess model performance and uncover disparities across demographic groups.

The primary focus of this research can be condensed into the following research questions, which will be elaborated on in subsequent chapters and that will guide the experiments.

"To what extent does race and gender bias exist in the zero-shot retrieval task involving text and image modalities in vision-language models? Are specific demographic groups consistently misrepresented or underrepresented in this scenario?"

"Is there evidence of race and gender bias in zero-shot classification tasks performed by vision-language models, and if so, how does this bias manifest across different demographic groups? Are specific demographic groups consistently misclassified in this setting?"

By examining the complexities of bias in vision and language models, this thesis aims to contribute to the ongoing research surrounding fair artificial intelligence advancements. Through analysis and empirical findings, the aim is to emphasize the significance of addressing biases in this field, particularly within these models, given their recent and widespread adoption which is expected to increase in the coming years, and also underline the importance of giving them the necessary consideration.

In the following chapters, the details of the methodology will be explored, the findings of the experiments will be unveiled and reflected upon, and interpretations will be provided to enrich the comprehension and understanding of bias in vision and language models.

2 Vision-Language Models and Bias

Vision and Language models, or VL models (VLMs), are advanced artificial intelligence systems that combine the understanding of visual and textual information. One characteristic that helps define these models is in fact their ability to process both vision and language - the first regarding images or videos, and the latter regarding natural language-based data. Vision-language models are developed to blend elements of computer vision, natural language processing, and machine learning techniques, and their general functionalities depends on the inputs, outputs, and the task these models are asked to perform [4]. Vision models analyze visual data to recognize objects, scenes, patterns, and relationships. They can perform multiple tasks like image classification, image captioning, image segmentation, object detection, and much more. Language models, on the other hand, process and understand textual information using natural language processing and deep learning techniques. They can perform tasks such as text classification, sentiment analysis, machine translation, question-answering, text generation, and much more.

Thanks to recent developments and remarkable technological advancements, and also thanks to the number of growing potential applications, in recent years, vision and language models have gained increasing popularity, leading also to improvements in their capabilities. The significant advancements achieved by these models up to today are progressing rapidly and these technologies are beginning to be used in an increasing number of applications, inevitably leading to real consequences. Therefore, research interest is increasingly focused on investigating their security, fairness, biases, and discrimination. A prominent theme, often not given proper consideration, is related to investigating the biases that are inherent in these models. Biases can originate from various factors and come into play at different occasions, from the initial stages all the way through to the end of the entire process. The presence of biases in these models represents a recognized issue in both the modalities involved - vision and language - and should by no means be overlooked or not properly considered. These aspects have

already been taken into account by some research studies if we consider the two modalities individually, but not as much if we consider both together. This is still evolving, however, when it comes to the multimodal version and the combined applications of vision and language. Combining these two modalities, in fact, runs the risk of merging certain individual problematic aspects or generating new ones.

Multimodal learning is the broad term that refers to the field of machine learning where information from multiple modalities, such as text, images, audio, or other types of data, is combined and integrated to improve the performance of a model on various tasks. Basically, it refers to the practice of utilizing a single model to learn representations from various modalities, where different statistical features define different modalities. Multimodal models have gained significant popularity due to their ability to leverage information present in different modalities and perform tasks that require understanding and generation of content involving multiple data types [85] [139] [159]. In the context of vision-language models, multimodal learning refers to the development of models capable of comprehending and generating content from both visual and textual inputs, these models are trained to process and generate information that integrates images and language.

2.1 Components of Vision-Language Models

A vision-language model typically incorporates three fundamental components: an image encoder, a text encoder, and a strategy to fuse information from the two encoders [4]. The visual encoder processes visual inputs and encodes them into a fixed-dimensional representation; it is responsible for processing visual information - such as images or video frames - and transforming it into a meaningful representation. It extracts high-level features from the visual data, capturing information about objects, scenes, and visual context; this extracted features can then be used in conjunction with the data captured by the language encoder for various tasks. To learn image features, some architectures have been widely adopted: CNN-based architectures (Convolutional Neural

Networks) [77], and Transformer-based architectures [124] in particular. The language encoder handles textual inputs and encodes them into a fixed-dimensional representation; it captures the semantic and syntactic information from the text, and its job is to process textual information and convert it into a meaningful numerical representation that can be combined with visual features for various multimodal tasks. It can be based on Recurrent Neural Networks (RNNs) [108], or more advanced models like Transformers and its variants [124]. The advent of this technology has radically changed the architecture and structure of these models.

The latest research, in fact, predominantly adopts image and text encoders with Transformer architectures to separately or jointly learn image and text features [33]. The Transformer architecture, originally introduced for Natural Language Processing (NLP) in 2017 [124], revolutionized the way sequence-to-sequence tasks were approached by introducing a self-attention mechanism that, in the context of VL models, enables the models to capture interactions between visual and textual elements, allowing for a more comprehensive understanding of the context. The Transformer architecture provides a powerful framework for processing both visual and textual modalities, and it is the overarching framework that incorporates attention mechanisms and cross-modal embeddings to enable the fusion, alignment, and processing of visual and textual information in vision-language models. Transformers have replaced RNNs as the standard model for most NLP tasks, with pretrained transformer models such as GPT [14] [92] as state-of-the-art [144], and, while CNNs are still widely used, they can be also adapted to CV tasks with competitive results [33]. Because they can achieve performance levels close to the state-of-the-art in both domains, transformers have emerged as the obvious choice as the foundation for pre-trained vision and language models. At a high-level, this architecture follows an encoder-decoder structure: the encoder on the left-hand side of the architecture extracts features from an input sequence, while the decoder on the right side uses those features to produce the output sequence.

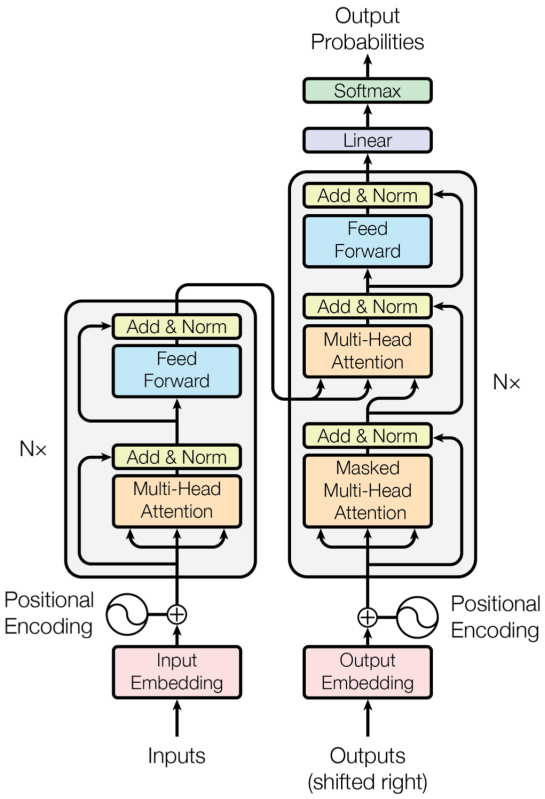


Figure 1: The Transformer Model [124]

In the Transformer architecture [33], the encoder - that consists of N layers - maps an input sequence ($x = (x_1, ..., x_n)$) into an intermediate representation ($z = (z_1, ..., z_n)$) using multiple layers, each consisting of a Multi-Head Attention (MHA) sub-layer and a Feed-Forward Network sub-layer (FFN). The MHA sub-layer enables the model to capture dependencies across the input sequence, while the FFN sub-layer introduces non-linearity. Considering z as input, the decoder generates an output sequence $y = (y_1, ..., y_n)$. Residual connections and layer normalization are then applied to each sub-layer. The decoder, which also consists of N layers, includes a third MHA sub-layer that attends to the encoder's output. It uses a modified self-attention mechanism to mask previous positions in the sequence and generate predictions based on prior outputs, making it suitable for generative tasks. This multi-head attention mechanism remains central to the transformer's functionality [33] [124], enabling effective modeling of long-range dependencies and contextual relationships. Since the attention mecha-

nism doesn't have an inherent order, in addition to the mentioned components the Transformer architecture incorporates positional encoding, which is added to the input embeddings to provide positional information, allowing the model to understand the sequential order of tokens within the input sequences.

The outlined structure describes the original encoder-decoder architecture for Transformers. However, it is important to mention the existence of more Transformer architectures. These include encoder-only models like BERT (Bidirectional Encoder Representations from Transformers) [29] and RoBERTa [81], and decoder-only architectures like the GPT (Generative Pre-trained Transformer) Series [100] [14] [92]. Models that employ the encoder-only architecture prioritize tasks such as contextualized word embeddings and bidirectional language understanding, enhancing comprehension of sentence semantics and syntactic structures. This architecture mitigates the limitations of lacking contextual understanding in traditional models like Bag-of-Words (BoW) and basic word embeddings. Encoder-only transformers address this limitation by considering the entire sentence simultaneously through multi-head self-attention, bidirectional understanding, and positional embeddings. Key components include multiple encoder layers, positional embeddings, and the self-attention mechanism. Conversely, decoder-only Transformer architectures excel in generating coherent and contextually relevant text. These models generate text sequentially, considering previous context and ensuring thematic consistency. Key components include multiple decoder layers with masked multi-head self-attention, encoder-decoder attention, and positional embeddings.

2.2 Model Architectures

Vision-language models need a model architecture that enables interaction between the features from textual and visual modalities [33]. As previously mentioned, VLMs incorporates three fundamental elements - an image encoder, a text encoder, and a strategy to fuse information from the two encoders -, and the latest research primarily relies on

the Transformer architecture for this purpose. Different model design choices can be based on pre-trained VL transformers, and an important distinction lies in how textual and visual modalities interact [33] [34]. We can categorize VLMs based on whether this interaction is shallow or integrated within the deep learning model itself. Models that rely on shallow interaction are referred to as dual encoders, while models using deep interaction are referred to as fusion encoders, which can employ different architectural designs.

Dual encoders processes visual and textual representations independently - images and texts are in fact encoded separately -, without interaction occurring within the deep learning model. Instead, they use a simple mechanism, like cosine similarity, for the visual and textual module outputs to interact. In fusion encoders, however, additional Transformer layers are added in the architecture to model the deep interaction between image and text representations, besides the use of an image encoder and a text encoder [7] [34].

Fusion encoders can be classified into two main categories: single-tower (one-tower) and dual-tower (two-tower) architectures [17] [7] [33]. In a single-tower architecture, a single transformer encoder operates on a concatenation of visual and textual input representations. Given that both visual and textual tokens are embedded into a single input, the single transformer stack facilitates unrestricted modeling of modality interactions. While single-tower models may exhibit variations in other aspects, such as embedding strategy, pre-training tasks, and data sources, the core architecture remains largely consistent across all single-tower models. Additionally, it offers the benefit of needing fewer parameters compared to the more intricate two-tower architecture. Rather than operating on a concatenation of visual and textual inputs, two-tower architectures encode each modality in separate transformer stacks, and interaction is then achieved through a cross attention mechanism.

A number of recently introduced VL models aim to leverage the advantages of both dual encoders and fusion encoders within a single model, resulting in combination encoders [33]. These models contain separate visual and textual encoders at the base of the model, and, before entering a fusion encoder module of various designs, the outputs of the text encoder and image encoder are aligned using cosine similarity.

Based on the original Transformer architecture, some VL models choose a structure that involves at least one encoder stack and one decoder stack. This architectural choice, known as encoder-decoder, is known for its versatility and allows models that employ it to effectively handle various tasks, including generative tasks [33] [34] [7].

In addition, other studies divide VLP methods based on fusion encoder between two-stage pre-training pipeline and end-to-end pre-training methods [34]. The first method, adopted in early researches, is characterized by the fact that image region features are first extracted from a pre-trained object detector. In end-to-end VLP methods, that are becoming more popular, image features are extracted from either Convolutional Neural Networks, Vision Transformers, or only using image patch embeddings, and the model gradients can be back-propagated into the vision backbone for end-to-end training.

2.3 Pre-Training Objectives

Various vision-language pre-training objectives have been designed for learning rich vision-language correlations. These objectives guide the learning process of the models by defining what they should aim to achieve during the training phase, so each objective represents a particular way of teaching the model to understand and generate connections between visual and textual information. They can be broadly grouped into three categories: contrastive objectives, generative objectives, and alignment objectives [153] [152].

Contrastive learning has been widely investigated in VLMs pre-training, which designs contrastive objectives for learning discriminative image-text features. Contrastive objectives train vision-language models to learn discriminative representations by pulling paired samples close and pushing others faraway in the feature space [47] [141]. Three main types of contrastive learning objectives can be distinguished: image contrastive learning focuses on learning discriminative features in images, image-text contrastive learning aims to learn the correlation between images and text by contrasting pairs of them, and image-text-label contrastive learning introduces image classification labels into the image-text contrastive learning process, allowing for the simultaneous learning of discriminative and task-specific features [33] [153]. Other common techniques are siamese networks [76] and triplet loss [94]. In the case of siamese networks, these architectures consist of two identical subnetworks with shared architecture and weights; each subnetwork is responsible for processing one modality, such as an image or a text description. The primary objective during training is to ensure that siamese networks produce embeddings in such a way that the L2 (Euclidean) distance between matching pairs is minimized, while the distance between non-matching pairs is maximized. Triplet loss introduces a different approach, which focuses on triplets of data points: an anchor (for instance an image), a positive sample (a matching image or text), and a negative sample (a non-matching image or text). The goal is to minimize the distance between the anchor and the positive sample while maximizing the distance between the anchor and the negative sample.

Generative objectives learn semantic features by training networks to generate image/text data via image generation, language generation, or cross-modal generation [153]. Four approaches can be distinguished: in masked image modeling, certain patches in an image are masked, and the model is trained to reconstruct them based on the unmasked patches; masked language modeling involves masking a fraction of tokens in text inputs and training the network to predict the masked tokens; masked cross-modal modeling combines both image patches and text tokens by masking and reconstructing

them jointly; while image-to-text generation focuses on generating descriptive texts for a given image, aiming to capture fine-grained vision-language correlations [33] [153].

Alignment objectives align the image-text pair via global image-text matching or local region-word matching on embedding space [153]. Two types of matching objectives in the context of VLMs are image-text matching, that focuses on establishing global image-text correlation by directly aligning paired images and texts, and region-word matching, that aim to model local, fine-grained vision-language correlations by aligning image regions with word tokens [153] [33].

In brief, VLM pre-training employs various cross-modal objectives to model vision-language correlation, including image-text contrastive learning, masked cross-modal modeling, image-to-text generation, and image-text/region-word matching. Additionally, unimodal objectives can be explored for maximizing the potential of each modality, such as masked image modeling for images and masked language modeling for text. Recent VLM pre-training emphasizes learning global vision-language correlation, benefiting image-level recognition tasks like image classification. Furthermore, some studies focus on modeling local, fine-grained vision-language correlation through region-word matching to improve dense predictions in tasks like object detection and semantic segmentation.

2.4 Pre-Training Methods

Vision-language models are typically pre-trained on large-scale datasets that contain visual and textual data in a paired way. This pre-training phase involves training the model to learn general representations of both modalities before fine-tuning it on specific downstream tasks. Fine-tuning allows the model to adapt its capabilities to desired applications. In recent years, however, the goal to reach for larger VLMs is to work in zero or few-shot settings and perform in-context-learning, that corresponds to an

emergent behavior in Large Language Models where the LLM performs a task just by conditioning on input-output examples, without optimizing any parameters [14] [136].

Unimodal pre-training involves learning language or visual representations separately from textual or image data, while multimodal pre-training links representations between images and textual descriptions [21] [31]. VLM pre-training works with a deep neural network that extracts image and text features from N image-text pairs within a pre-training dataset $D = \{x_n^I, x_n^T\}_{n=1}^N$, where x_n^I and x_n^T denote an image sample and its paired text sample. The deep neural network has an image encoder f_θ and a text encoder f_Φ, which encode the image and text (from an image-text pair $\{x_n^I, x_n^T\}$) into an image embedding $z_n^I = f_\theta(x_n^I)$ and a text embedding $z_n^T = f_\Phi(x_n^T)$, respectively [152].

2.5 Attention Mechanisms and Cross-Modal Embeddings

Attention mechanisms enable vision-language models to focus on relevant parts of input sequences or images, enhancing their understanding and generating more accurate outputs, and are employed to capture the relationships between visual and textual information. These mechanisms calculate attention scores that represent the relevance or similarity between each visual feature and each textual embedding - the attention scores are then used to weight and combine the visual features and textual embeddings, creating a fused representation that combines both modalities [40]. Various attention mechanisms can be categorized according to their approach [41]. From the recognition process of the human visual system, we can extract a general form - or a general approach - for attention mechanisms: when seeing a scene in our daily life, in fact, our focus concentrate on some discriminative regions, and it allows us to process these regions quickly. The procedure described above can be formulated as follows: $\mathit{Attention} = f(g(x), x)$, where $g(x)$ can represent to generate attention which corresponds to the process of attending to the discriminative regions, $f(g(x), x)$ means processing input x based on the attention $g(x)$ which is consistent with processing crit-

ical regions and getting information. There are various attention mechanisms, such as channel attention, spatial attention, temporal attention, and branch attention, but almost all the existing ones can be written into the general formula specified above [41]. Channel attention, where different channels in different feature maps usually represent different objects, adaptively recalibrates the weight of each channel, and it can be viewed as an object selection process, thus determining what to pay attention to. Spatial attention, instead, can be seen as an adaptive spatial region selection mechanism: where to pay attention. Temporal attention can be viewed as a dynamic time selection mechanism determining when to pay attention, and is thus usually used for video processing, while branch attention can be seen as a dynamic branch selection mechanism: which to pay attention to, used with a multi-branch structure. Other than the categories described above, it is possible to combine the advantages of two of them, like in channel & spatial, that adaptively selects both important objects and regions, and spatial & temporal attention, that adaptively selects both important regions and key frames. As previously stated, in VL tasks the core of the attention mechanism is the alignment score calculation, that quantifies the amount of attention that the visual features would place on each of the language representations - or linguistic features would empower on the specific visual regions - when bridging the semantic gap between visual and language features [19]. Regarding multimodal VL downstream tasks, two commonly used approaches are the cross-attention, that is performed between visual and textual inputs, and the self-attention, that is performed over all inputs within each modality [19].

Cross-modal embeddings bridge the gap between visual and textual information by mapping them into a shared space, facilitating multimodal understanding. VL models are trained to align visual and textual data to achieve cross-modal embeddings [34] [152] [88]. These embeddings are learned through joint training on large-scale datasets: during this training, the model learns to map visual features and textual embeddings into a shared embedding space - this shared space allows for direct comparison and

alignment between visual and textual representations. The learning process involves optimizing the model's parameters to ensure that when an image and its associated text are projected into the shared embedding space, they are close to each other. This alignment can be learned through various techniques in which the model is trained to bring matching pairs of data closer in the embedding space and push non-matching pairs apart.

As explained above, textual and visual input must be encoded into a sequence of - respectively - textual tokens $t_1, ..., t_T$ and visual features $v_1, ..., v_V$; where each element in the sequence is a numerical vector [33]. While the majority of VL models employ the same embedding strategy for textual representations, the strategy for representing images varies significantly and stands as one of the key differences in pre-trained VLMs. The strategies to handle textual representations utilizes the textual embedding approach of BERT, using the WordPiece algorithm for tokenization, starting with a '[CLS]' token, and separating text sequences with '[SEP]' tokens [29]. They also utilize learned embeddings for token position and segment identification to create the token's input representation. Some models, like CLIP [101], use BPE encoding, while others such as BEiT-3 [130] and Flamingo [5] opt for the SentencePiece encoding method. Despite these variations, the fundamental embedding strategies in these models remain quite similar. On the other hand, VLMs use various strategies to handle visual embeddings, such as region features - that employ region-based features generated by object detection networks -, grid features - extracted from a CNN's feature grid output -, and patch embeddings - that involves dividing images into patches and then projecting them into an embedding space while integrating position and type information [33].

2.6 Tasks Tackled by Vision-Language Models

The field of vision, language, and vision-language research, as already expressed above, is facing rapid changes in trends, and the progresses are fast-paced. For this reason, it's important to identify the current developments and trends regarding the tasks that

form the foundation of current multimodal VL research. The tasks tackled by VL models can be divided in three main categories: generation tasks, classification tasks, and retrieval tasks [123] [88].

Generation tasks include visual question answering (VQA) [3], that refers to the process of correctly providing an answer to a question given a visual input (image or video), and visual captioning (VC) [139], that involve generating syntactically and semantically appropriate descriptions for a given visual input. To generate relevant captions, rich linguistic knowledge and coherent understanding of the visual input is required. Other tasks that belongs in this category are visual commonsense reasoning (VCR) [149], that infers common-sense information and cognitive understanding given a visual input, and visual generation (VG) [102], that is concerned with generating visual output from a given textual input. Natural language for visual reasoning (NLVR) [50] - that determines if a statement regarding a visual input is correct or not - is a subtask of the broader category of VCR confining to the classification paradigm. Multimodal affective computing (MAC) is the prominent task of classification tasks. MAC interprets visual affective activity from visual and textual input, so the automated recognition of affective phenomenon causing or arising from emotions. Visual retrieval (VR) is the core task of retrieval tasks. It corresponds to the task of retrieving images based only on a textual description. Text-image retrieval is a cross-modal task involving the understanding of both language and vision domains with appropriate matching strategies, and its aim is to fetch the top-most relevant visuals from a larger pool of visuals as per the text description. Other tasks can be identified that are not included in the previous differentiation: vision-language navigation (VLN) [39] and multimodal machine translation (MMT) [143]. The first is the task of an agent navigating through a space based on textual instructions, while the second is a two-fold task that involves translating a description from one language to another with additional information from other modalities, such as video or audio. Apart from the distinctions above, VL tasks can be also differentiated based on other characteristics, such as if their output is textual or

not [123].

2.7 State-of-the-Art Vision-Language Models

In recent years, several vision and language models have been developed. Among these, the models defined as *state-of-the-art* (SOTA) are the most advanced and highly performing models in this field of study within a confined research period. These models represent the cutting-edge of research and development, and they often achieve the highest levels of performance on specific tasks or benchmarks. It is certainly essential to consider the speed at which different models enhance their performance in this field of study, as well as how advancements in research are giving rise to new and powerful models on a constant basis. In the table below (Table 1) we can observe a list of representative vision-language pre-trained models, along with their year of publishing and their architecture type.

Model	**Year**	**Architecture**
ALBEF [70]	2021	Combination Encoders
ALIGN [60]	2021	Dual Encoder
BeiT-3 [130]	2023	Combination Encoders
BLIP-2 [71]	2023	Encoder-Decoder
BridgeTower [140]	2022	Two-Tower Encoder
BriVL [55]	2021	Two-Tower Encoder
CLIP [101]	2021	Dual Encoder
CoCa [145]	2022	Encoder-Decoder
DaVinci [30]	2022	Encoder-Decoder
E2E-VLP [137]	2021	Encoder-Decoder
Flamingo [5]	2022	Encoder-Decoder
FLAVA [114]	2021	Combination Encoders
Florence [147]	2021	Dual Encoder
GPV [43]	2021	Encoder-Decoder
LEMON [52]	2021	Encoder-Decoder
LXMERT [120]	2019	Two-Tower Encoder
mPLUG [66]	2022	Encoder-Decoder
OFA [129]	2022	Encoder-Decoder
OmniVL [127]	2022	Encoder-Decoder
OneR [58]	2022	One-Tower Encoder
OSCAR [75]	2020	One-Tower Encoder
PaLI [22]	2022	Encoder-Decoder
Pixel-BERT [53]	2020	One-Tower Encoder
SimVLM [131]	2021	Encoder-Decoder
SOHO [54]	2021	One-Tower Encoder
UNIMO [74]	2020	One-Tower Encoder
UniTAB [142]	2022	Encoder-Decoder
UNITER [24]	2019	One-Tower Encoder
ViLBERT [82]	2019	Two-Tower Encoder
ViLT [62]	2021	One-Tower Encoder
VinVL [154]	2021	One-Tower Encoder
VisualBERT [73]	2019	One-Tower Encoder
VL-BERT [118]	2019	One-Tower Encoder
VL-T5 [25]	2021	Encoder-Decoder
VLMo [7]	2021	Combination Encoders
X^2-VLM [150]	2022	Combination Encoders

Table 1: List of representative *State-of-the-Art* Vision and Language Models

2.8 Multimodal Data

Due to the high cost of human annotation, there are only limited high-quality human-annotated image-text datasets pairs. A major part of recent works relies heavily on noisy data extracted from the web, exploiting the technique of web scraping. The web scraping process is in fact the fastest way to collect a large number of image-text pairs without the need for humans in the loop. However, as the data is noisy, wrong description of an image could affects the quality of training of vision-language models [72]. As stated above, VL models are usually pre-trained on large-scale datasets that contain visual and textual data in a paired way, to learn general representations of both modalities before adapting them to specific tasks. Some datasets are more popular and have been used as the foundation for the pre-training phase of several models, whereas other models have been trained on datasets specifically created for their needs.
This section aims to summarize some commonly used large-scale datasets for vision-language models pre-training and evaluation, along with the year of publishing, the number of image-text pairs, and the language.

Dataset	**Year**	**Image-Text Pairs**	**Language**
SBU Caption [93]	2011	1M	English
COCO Caption [23]	2015	1.5M	English
Yahoo Flickr CC 100M (YFCC100M) [121]	2016	100M	English
Visual Genome (VG) [64]	2017	5.4M	English
Conceptual Captions (CC3M) [112]	2018	3.3M	English
Localized Narratives (LN) [99]	2020	0.87M	English
Conceptual 12M (CC12M) [20]	2021	12M	English
Wikipedia-based Image Text (WIT) [116]	2021	37.6M	100+
Red Caps (RC) [28]	2021	12M	English
LAION400M [109]	2021	400M	English
LAION5B [110]	2022	5B	100+

Table 2: Popular Image-Text Datasets

Other than the ones printed out in the table above, other datasets are used to train the current state-of-the-art vision-language models. Between those we can find the datasets used to train the following models, all collected from the web and not public: CLIP [101], that contains 400M image-text pairs in English language, ALIGN [60], with 1.8B

image-text pairs also in English language, and WebLI [22], that contains 12B pairs in over 100 languages.

2.9 Bias in Vision-Language Models

According to the National Institutes of Health [95], bias is defined as any tendency which prevents unprejudiced consideration of a question. Bias, in fact, refers to the systematic and consistent deviation from a true or objective representation of a particular phenomenon, group, or set of data, resulting from various factors, including cognitive, social, cultural, or methodological influences. Biases can also be defined as human prejudice [18], they can occur at any phase of research, and they can exist in many shapes and forms. Biases hold significant social implications, and when viewed from the modeling perspective they can have profound effects on the robustness and generalization of - in this case - VL models.

Multimodal models like vision-language models tend to exhibit so-called 'human-like' biases [18], in particular stereotypical biases regarding gender and race, but also biases about age, disability, nationality, ethnicity, and more. In recent years, the academical studies regarding this topic have started to grow, as these problems are receiving more attention in society. The more these models are used in practical applications, in fact, the more discriminating they are toward the afflicted population groups and the more detrimental they can be to society. Given that the possibly affected groups collectively make up a large portion of the global population, this is concerning. Investigating the problem of multidimensional bias diffusion and its impact in vision-language models is a relevant topic to consider to develop fair and - in general - better models.

The majority of AI systems and algorithms rely on data and require it for training. Data and the functioning of these methods and systems are intricately connected. If biases exist in the training data, the algorithms trained on that data will learn and integrate these biases into their predictions. Consequently, pre-existing biases in the data can influence the algorithms that utilize it, resulting in biased outcomes. Algorithms

can also amplify and perpetuate existing biases in the data [86]. Without the right adjustments, we can enter in a never-ending loop that goes from bias in the data to bias in the models until it generates bias in the decision making. Therefore, before analyzing the datasets designed to address bias in VL models, and the academical studies in literature, it is important to understand the concept of bias.

A classical example of bias consists of the so-called 'sample bias', but there are many possible categorizations within this subject. They can be grouped according to biases in data, which result in biased algorithmic outcomes, or biases in user behavior, which are a result of algorithmic outcomes and affect user behavior as a consequence, or even biases in users that might be reflected in the data they generate [86].
The sampling bias lead the way of the first category regarding bias in data. It arises due to non-random sampling of subgroups, and as a consequence the trends estimated for one population may not generalize well to data collected from a new population. Other relevant biases in data [119] are measurement bias, that arises from how we choose, utilize, and measure particular features, omitted variable bias, that occurs when one or more important variables are left out of the model, representation bias, that arises from how we sample from a population during data collection process, and aggregation bias, that occurs when false conclusions are drawn about individuals from observing the entire population [86].
Another distinction regards the category of bias that arises from human behavior [6] [91]. The algorithmic bias occurs when the bias is not present in the input data and is added purely by the algorithm, the user interaction bias is a type of bias that can be observant on the web but also get triggered from two sources like the user interface and through the user itself by imposing his/her self-selected biased behavior and interaction. But also the popularity bias, where items that are more popular tend to be exposed more, the emergent bias, that occurs as a result of use and interaction with real users, and the evaluation bias, that happens during model evaluation.
To conclude this section about bias, we can list the bias that users might reflect in

the data they generate [86] [6] [91]. The historical bias is the pre-existing bias and socio-technical issues in the world that can infiltrate the data generation process, even with perfect sampling and feature selection. The population bias arises when statistics, demographics, representatives, and user characteristics are different in the user population of the platform from the original target population, whereas the social bias happens when others' actions affect our judgment. The behavioral bias arises from different user behavior across platforms, contexts, or different datasets, while the temporal bias arises from differences in populations and behaviors over time.

Another relevant topic linked to bias is algorithmic fairness. There are various definition of fairness, but in this context we can define it as the absence of any prejudice or favoritism towards an individual or a group based on their intrinsic or acquired traits in the context of decision-making [106]. There have been also numerous attempts to address bias in artificial intelligence in order to achieve fairness, from pre-processing to post-processing [86], that consequently lead to many options to unbias data and models.

To this day, the topic related to bias in vision-language models is still understudied, especially in its multidimensional entirety, and it struggles to keep up with the developments of new models. In fact, most studies conducted so far consider vision and language individually. Regarding bias in Natural Language Processing (NLP), statistical patterns of human language are reflected by large-scale language models, which might be problematic if training datasets include offensive or inaccurate words [132], and previous studies have manifested the presence of racial bias [36] [84], as well as gender bias [157] [12]. Similar discourse in the case of Computer Vision, with investigations that have also shown evidence of gender bias [158], and racial bias [133]. However, while in an initial phase biases were assessed mainly on language models and vision models separately, recently a part of literature started to investigate bias in multimodal models. Considering the aim of this work, this is the specific area we will focus on. It is still important to address the findings on this topic individually, since bias can be generated

from both vision and language, with different intensity, or from only one of the two. To conclude, it is important to note the realization of some debiasing techniques, designed to reduce the presence of bias in VL models, mainly related to specific tasks.

In literature, bias has been demonstrated to perpetuate both from vision and language [105]. The major part of work on language models focuses on gender and racial bias assessment [42] [13], but there have been also some studies that considered bias with regard to more categories like religion, disability, and profession [89] [56], and also linguistic bias - bias referred to the language itself [146] [90]. As other studies pointed out [117] [103], these types of bias are not exclusive to the language domain: in fact, image classifiers as well as multimodal models have also been shown to incorporate them. Regarding vision, Wang et al. [126] assessed the gender and racial bias in CLIP's image classification module using the Fairface dataset, and they further demonstrates that CLIP more strongly correlates male-gendered words with high-paying occupations than female-gendered phrases [125]. Additional work offer insights into possible CLIP model uses, as well as future research and evaluation of its gender and racial bias and assessing the differences in misclassification between various groupings [2]. Bhargava et al [9] measured gender inequality in several picture captioning systems, and proposed valid solutions, while another study revealed that zero-shot vision-language models, like CLIP, show gender-based performance disparities for many visual concepts [44].

Some more recent studies highlighted the presence of bias in vision-language models. Smith et al. [115] sustain the lack of validity of these models due to datasets' bias, and they evaluated the performances of CLIP and its variants with respect to gender, proposing a pipeline to balance the datasets. Berg et al. [8] investigated bias metrics to quantify racial and gender bias in facial images; they proposed ranking metrics, demonstrating that they are effective bias measures. They also added a supervised adversarial debiasing method that, based on the used metrics, showed a substantial reduction. Some results illustrate in Seth et al. [111] show how the original CLIP model focuses on the face and hair region for certain occupation-related keywords, and how, after debiasing, the focus shifts to more indicative cues (like the stethoscope for

a doctor). Again, the CLIP model exhibits a bias omitting the woman for the 'doctor' keyword, and only detecting the female medical professional as a 'nurse'. They also designed DEAR, a simple and effective debiasing method for VLMs that only adapts the image encoder of the VLM by adding a learned residual representation to it. Race, gender, and age biases receive significant attention in society and - as a consequence - also from researchers of this field of study, so they have been observed while analyzing various tasks. Other works focused instead on more specific types of bias, like sexual objectification [134]; sexual orientation, nationality, and disabilities [59] - or even sports-related gender bias [46].

Several multimodal vision-language models - like ALBEF, VisualBERT, and CLIP - are examined by Zhou et al. [161] to measure how often these models choose stereotyped statements to caption images that defy stereotypes. Garcia et al. [35] conducted a comprehensive analysis of the annotations of the Conceptual Captions dataset [112] - named PHASE (Perceived Human Annotations for Social Evaluation) -, focusing on how different demographic groups are represented. PHASE attributes are age, gender, skin-tone, ethnicity, emotion, and activity. They evaluated three prevailing vision-and-language tasks: image captioning, text-image CLIP embeddings, and text-to-image generation, showing that societal bias is a persistent problem in all of them. Zhang et al. [155] started by constructing a dataset of counterfactual template-based image-text pairs for measuring gender bias in pre-trained VL models. Later, they compared the difference between masked prediction probabilities of factual and counterfactual examples. Wang et al. [128] propose FairCLIP to eliminate the social bias in CLIP-based image retrieval tasks, a method that has the best compatibility between the debiasing effect and retrieval performance and can be effective in other tasks based on CLIP. Additional works focused on prompts to debias VLMs in tasks like text-image retrieval and image generation [26], while Bianchi et al. [10] demonstrated the extent of stereotypes and complex biases present in image generation models and the images generated by them - they show that simple user prompts can generate thousands of images that perpetuate dangerous stereotypes based on race, ethnicity, gender, class, and intersectionality [105].

Below, instead, we can delve deeper into some interesting datasets, with a lower number of image-text pairs but specifically used to investigate bias in VLMs. Some of these datasets are considered part of the computer vision category, but given the fact that they play an important role in vision-language research, and considering their relevance in studies exploring bias, their inclusion in this context ensures completeness. FairFace [61] consists of a set of cropped face images of different persons; it has over 100K images, collected from the YFCC-100M Flickr dataset [121] and annotated with the binary gender (male and female), ethnic-racial classes (White, Black, Indian, East Asian, Southeast Asian, Middle Eastern, and Latino), and multiple age brackets. Another large-scale face dataset is UTKFace [156], that consists of over 20K face images with annotations of age, gender, and ethnicity (White, Black, Asian, Indian, and Others - like Hispanic, Latino, Middle Eastern). CelebA - CelebFaces Attributes Dataset - is a large-scale face attributes dataset with more than 200K celebrity images, each with 40 attribute annotations. The images in this dataset cover large pose variations and background clutter, and the 40 binary labels indicate facial attributes like hair color, gender and age. VLStereoSet [161] consists of images divided by category (gender, profession, race, religion). Each image is accompanied by three candidate captions (taken from StereoSet [89]): one caption is stereotypical, one is anti-stereotypical and the third is semantically meaningless. One of these captions is labeled as the correct caption for the image, and the probing task is to identify this correct caption given the image. The EMOTIC dataset [63] focuses on the recognition of emotions in images. It consists of over 23K images, annotated with diverse emotional attributes: the images are in fact described with an extended list of 26 emotion categories combined with the three common continuous dimensions (valence, arousal, and dominance). This dataset has been used to study biases related to emotions and affective states in many contexts [65] [98] [97]. Other datasets, specifically developed to investigate bias in VL models, are available today. An example is VL-BIAS [155], that contains images with 52 activities and 13 occupations with a total of 24K image-text pairs. The PATA [111]

- Protected Attribute Tag Association - dataset consists of a list of images organized as a set of scenes, and a set of captions applicable to each scene, organized according to specific protected attributes. It consist of nearly 5K public images crawled from the web, organized in 24 scenes each with between 100 and 400 images. It considers the binary gender, five ethno-racial groups (Black, Caucasian, East-Asian, Hispanic-Latino and Indian), and two age categories (young, old). Additionally, PATA provides visual context for a diverse human population in different scenarios with both positive and negative connotations. The SOBEM [134] (Sexual OBjectification and EMotion Database) dataset is the sole standardized and controlled picture database available and designed to study sexual objectification. It contains 28 standardized photographs each of ten Caucasian women. Four emotional states are included: Neutral, Angry, Sad, and Happy; the Angry, Sad, and Happy states include high-emotion (more clearly visible on the face) and low-emotion (emotion more subtle) images. Each emotional state includes two photographs (hair is tied behind the head and hair falls loose over the shoulders) of the subject in a non-objectified condition, and two photographs in an objectified condition; in the objectified condition, the female subject is photographed from the waist up wearing a black bra but no shirt, while in the non-objectified condition, the same subject is photographed from the waist up wearing a black shirt which covers the entire chest. Lastly, MMBIAS [59] is a dataset that contains 3,5K target images and 350 English phrases covering 14 population subgroups and corresponding to different target concepts. Each target category has 250 corresponding images obtained from the popular image uploading website Flickr. The dataset also contains 20 textual phrases related to each target concept, used for bias experiments in the textual domain. For religion, it includes the five major religions in the world today: Islam, Christianity, Judaism, Buddhism, and Hinduism; for the national origin, it includes images corresponding to the four nationalities: American (USA), Chinese, Arab (collectively), and Mexican - French and Italian are also included in the textual phrases in addition to the former. For disability, it contains images for two common forms of disability: physical disability, mental disability as well as people with no disability; and in addition, the

textual data includes phrases corresponding to visual disability and hearing disability as well. The two most common types of sexual orientations, homosexuality and heterosexuality, are also included in MMBias.

2.10 Limitations and Testing of Vision-Language Models

Existing vision-language methods have two major limitations [72]. From the model perspective, most existing pre-trained models are not flexible enough to adapt to a wide range of vision-language tasks. From the data perspective, most models pre-train on image and alt-text pairs that are automatically collected from the web, but web texts often do not accurately describe the visual content of the images, making them a noisy source of supervision [72].

Considering all the models reviewed in the previous paragraphs, several limitations arise that need to be taken into account. Not all models encounter all the limitations that will be listed, so the goal is to provide a general overview of the main limitations characterizing vision-language models. To put under stress these models in order to exploit their limitations, several foiling techniques and benchmarks have been developed by researchers [96] [113] [162].

Vision-language models often struggle with compositionality, which is the ability to understand and generate complex expressions by combining known parts [83] [148] [122]. They may have difficulty in reasoning about unseen combinations or handling increased complexity in the input. In addition, there might not be a clear correlation between the size of the training dataset or model size and the models' compositional reasoning abilities - simply scaling up the data or the model does not guarantee improved performance in compositionality tasks.

From another perspective, given their amazing scalability, VL models can improve even more and achieve higher scores, but they require an astronomical amount of training resources. CLIP [101], for instance, can distinguish complex image patterns but struggles with handwritten digit recognition tasks: the authors attribute this type of

misclassification to a lack of handwritten digits in the training datasets [101]. Many VL models, in fact, also have poor generalization to images not covered in their pre-training dataset. VL models may experience degraded retrieval performances when faced with increased caption complexity: as the complexity of expressions or queries grows, the models may struggle to accurately retrieve relevant visual content [151]. Prompt engineering requires domain expertise and a user needs to spend a significant amount of time on words tuning since a slight change in wording could have a huge impact on performance [38] [160]. Pre-trained on large-scale image-caption datasets, VL models may be biased towards the patterns and distributions present in those datasets. This can result in limited generalization to new or diverse data, affecting their performance on real-world scenarios [162]. Some models struggle on more abstract or systematic tasks such as counting the number of objects in an image and on more complex tasks such as predicting how close an object is in an image - so computing the distance [78]. VL models may also have limited contextual understanding: they may struggle with understanding the contextual nuances and dependencies within a scene or a sentence, and they might not capture the subtle relationships between objects, attributes, and their interactions accurately, resulting in limited contextual understanding [16] [11] [107]. Zero-shot models like CLIP struggles compared to task specific models on very fine-grained classification, such as telling the difference between car models, variants of aircraft, or flower species. Some models, like BLIP-2 [71], use Large Language Models, and address the problems faced by LLM, such as outputting offensive language, propagating social bias, or leaking private information. A further problem is inaccurate knowledge from the LLM, activating the incorrect reasoning path, or not having up-to-date information about new image content. In addition, models like PaLI [22] offers multi-language use, but some of the multilingual capability is lost when the model is fine-tuned for English-only data. A multilingual multimodal model - trained on diverse datasets that encompass various languages and modalities - refers to an artificial intelligence system capable of processing and understanding multiple languages as well as multiple modalities of input, such as text, images, and audio [15] [80].

Vision-language models testing can be differentiated between multiple options. From dataset to model analysis, so from the beginning of the process to the end of it, there are various choices to investigate and put these models under stress, to exploit their bias and limitations. In the following paragraph, the most common testing techniques will be briefly examined to investigate their objectives and the processes involved. Dataset analysis starts by examining the datasets used to train the vision-language models. It focuses on analyzing the data sources, data collection methodologies, and potential biases present in the data, while looking for imbalances in demographic representation, cultural biases, or stereotypes that might be reflected in the training data. Hirota et al [49], for instance, presented a study where they investigate gender and racial bias in five VQA datasets, demonstrating the presence of harmful samples denoting gender and racial stereotypes. Benchmark datasets refers to the several solutions that have been developed to evaluate biases in vision-language models. For example, the COCO Captions dataset [23] includes human-written image descriptions that may exhibit biases; by analyzing model performance on these datasets, researchers can identify and quantify biases present in the models' output. This is directly related to the issue highlighted by van Miltenburg et al. [87] regarding annotators - like in the COCO dataset - and their inferences that, while understandable, are not necessarily supported by the image in question. Bias probing tasks involve designing specific tests to assess biases in VL models. These tasks typically involve asking the model to perform certain predictions or generate text based on visual inputs; researchers then analyze the outputs for any indications of bias. Counterfactual evaluation involves modifying inputs to VL models and analyzing how these modifications affect the model's output. By systematically altering aspects like gender, race, or other protected attributes in the input, researchers can examine how biases propagate through the model and influence its responses [1] [155] [90]. Human evaluation is crucial to understanding biases in VLMs. Researchers can design studies where human annotators assess the models' outputs for biases. By comparing human judgments with model predictions, biases can be identified and mea-

sured. Model analysis involves conducting a detailed analysis of the model's attention mechanisms, internal representations, or activation patterns. By examining which parts of an image or text the model focuses on, researchers can identify any biased patterns in the model's decision-making process. The choice of metrics may vary depending on the specific research goals, biases being examined, and the nature of the dataset and task at hand.

3 Methodology

As mentioned earlier, the aim of this study is to thoroughly examine and understand bias in vision and language models. This involves analyzing their presence, identifying causes and implications, but also exploring how bias propagate. To achieve this, three pre-trained vision-language models, each with unique characteristics, will be considered. These models will be utilized to perform two main tasks - namely zero-shot classification and zero-shot retrieval - aimed at addressing two research questions. The models in question, in alphabetical order, are ALBEF, BLIP-2, and CLIP. Before outlining the research questions and the methodology used to address each one, a more in-depth analysis of each model will be provided. This examination follows the exploration conducted in the previous chapter, presenting a more detailed overview of the features of each model. But first, it is necessary to illustrate the image dataset that has been specifically created for the purpose of this work and that will be used in later experiments.

3.1 Data Compilation

Starting from the UTKFace dataset [156], which is a large-scale face dataset, a new and enhanced 'morphed' dataset has been constructed. The UTKFace dataset comprises over 20,000 face images with annotations of age, gender, and ethnicity. The images encompass a broad range of variations, including pose, facial expression, illumination, occlusion, resolution, and more. This dataset served as the foundation for collecting the set of images that specifically represent the focal dimensions of this study: gender and race.

Figure 2: Sample images from UTKFace original dataset

The labels contain information about the dimensions of each image. Specifically, the labels for each face image are embedded in the file name, formatted as follows: *[age]_[gender]_[race]_[date&time].jpg*. The meanings of each label are explained below, as outlined in the original paper [156].

- age: An integer ranging from 0 to 116, indicating the age.
- gender: Either 0 (male) or 1 (female).
- race: An integer from 0 to 4, representing White, Black, Asian, Indian, and Others (such as Hispanic, Latino, Middle Eastern).
- date&time: In the format "yyyymmddHHMMSSFFF," indicating the date and time the image was collected for the UTKFace dataset.

In order to balance the categories in the dataset, five images were randomly chosen for each combination of gender, race, and age. For gender, the dataset includes Female and Male, while for race, the dataset includes Asian, Black, Indian, and White. To create the morphed dataset, 200x200-cropped and aligned images were transformed based on a selected dimension (gender or race), to generate new images of the same person,

differentiated only by the chosen dimension. As specified above, race considerations included only White, Black, Asian, and Indian, not considering the category named Others in the original dataset.

Two different techniques were employed to modify images according to race and gender. Initially, the images from the original dataset were organized into 40 subfolders. This division, following Hovy et al.'s proposal [51], considered age ranges (1-14, 15-24, 25-54, 55-64, and 65+) and race groups mentioned above, as well as genders. At this point, each subfolder contained five random images from the original UTKFace dataset, with names reflecting the demographic combination, such as *'Asian_Female_1-14'* and *'Indian_Male_15-24'*.

Regarding the modification of images to other races, a Telegram bot based on the Pix2Pix model was utilized. In short, Pix2Pix [57] is a type of conditional Generative Adversarial Network (GAN) that, unlike traditional GANs, learns to map input images to output images using paired datasets. It involves a generator and discriminator network, where the generator creates realistic output images matching target images, and the discriminator distinguishes between real and generated images. Through adversarial training, the generator improves image generation quality. This bot was capable of altering each image to depict individuals of other respective races. The interface allowed users to select, after uploading an image and choosing the desired output race, from three different editing options ('strong', 'normal', and 'weak'). To achieve more pronounced modifications, the 'strong' option was chosen - for the entire process - as it yielded the best results at the front-end. Consequently, each uploaded photo was modified to depict individuals of the other three respective races, resulting in four versions of the same person, distinguished by physical characteristics.

Regarding the modification of images to the opposite gender of individuals, an app called 'FaceApp' [79] was used. In this way, starting from each individual in the subfolders, the same individual, but of the opposite gender, was generated.

Finally, the morphed dataset can be divided into two main pairs: morphed-gender and morphed-race. The morphed-gender dataset comprises a total of 400 face images,

including the selected 200 and their corresponding morphed images. Meanwhile, the morphed-race dataset consists of 800 images, featuring the same selected 200 and their corresponding morphed images representing other races.

Target Concept	**Target Values**
Race	Asian, Black, Indian, White
Gender	Female, Male

Table 3: Target Classes and Target Groups in the Morphed dataset

Below, we can observe some examples of images present in the morphed dataset; the original photos are on the left or in the upper corner, while the corresponding modified images are next to them. Regarding race, starting from left to right we can observe examples for Asian, Black, Indian, and White, in this order.

Figure 3: Sample images from the morphed version of the UTKFace dataset | Race

Figure 4: Sample images from the morphed version of the UTKFace dataset | Gender

In addition to the image dataset presented above, a number of bias words will be considered; these words will constitute the corresponding captions used as textual component in each task. Following a similar approach to the work of Berg et al. [8] and Wang et al. [128], pairwise adjectives are selected from the IAT dataset [138]. These pairs of

adjectives - opposite in meaning - refer to a set of contrasting attributes or characteristics commonly employed in implicit association tests (IATs) to evaluate implicit biases or preferences towards specific social groups.

In particular, only pairs of words uncorrelated with facial expressions or sensitive attributes are used. For these reasons, pairs like 'happy/sad' are not included, since it has correlations with facial expressions. Finally, twelve bias words, potentially containing human bias, are chosen. The caption template and bias words are detailed in the table below.

Caption Template	**Bias Words**
A photo of a {} person	smart, stupid, rich, poor, nice, terrible, kind, evil, lawful, criminal, trustworthy, dishonest

Table 4: The Caption Template and the Bias Words used in this work

The framework assumes there is a reasonably concrete normative view that there exists a set of 'neutral' text queries like "a smart/stupid person" which hypothetically should be independent of demographic categories [8]. In general, some associations with a sensitive attribute are legitimate and explainable, while others are illegitimate and unjust. The choice of these words, combined with the composition of these sentences, was done to create gender- and race-neutral queries. This ensures that the captions are not associated with any specific race or gender, providing a benchmark where it is possible to associate each image with all the captions from the textual template. These words - and sentences - will then be used in the following tasks, allowing for the assessment of the potential presence of bias.

3.2 Selected Vision-Language Models

In the previous chapter, we explored various vision-language models. Now, let's delve into the three models that will be involved in the upcoming tasks.

ALBEF Align before Fuse - ALBEF [70] - introduces a contrastive loss to align image

and text representations before fusing them through cross-modal attention, enabling more grounded vision and language representation learning.

ALBEF comprises an image encoder (ViT-B/16), a text encoder (first 6 layers of BERT), and a multimodal encoder (last 6 layers of BERT with additional cross-attention layers). The model tries to leverage the advantages of dual encoders and fusion encoders. It contains separate visual and textual encoders at the base of the model, and, before entering a fusion encoder module of various designs, the outputs of the text encoder and image encoder are aligned, resulting in an architecture called combination encoders, as seen in Table 1. The model employs three pre-training objectives: image-text contrastive learning (ITC), masked language modeling (MLM), and image-text matching (ITM). ITC is applied to the unimodal image encoder and text encoder, and it aligns the image features and the text features; ITM is applied to the multimodal encoder and it predicts the positivity of image-text pairs (matched or non-matched); MLM is also applied to the multimodal encoder and utilizes both the image and the contextual text to predict the masked words. Additionally, ALBEF proposes momentum distillation to learn from noisy data.

The pre-training data is constructed using two web datasets (Conceptual Captions, SBU Captions) and two in-domain datasets (COCO [23] and VisualGenome [64]). The total number of unique images is 4M, and the number of image-text pairs is 5.1M. To be scalable with larger-scale web data, the much noisier Conceptual 12M dataset [20] is included, increasing the total number of images to 14.1M.

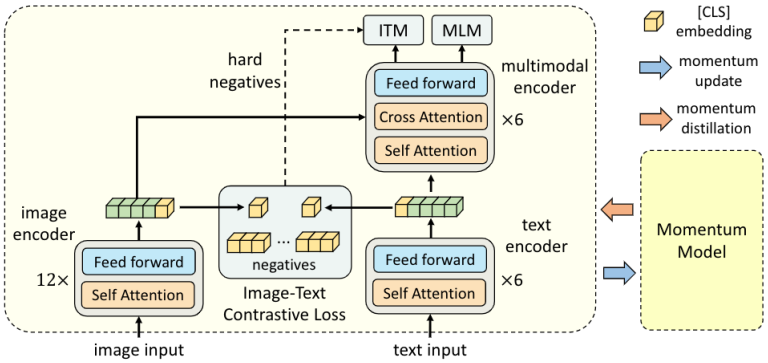


Figure 5: General Overview | Framework of ALBEF [70]

BLIP-2 Developed by Salesforce, BLIP-2 [71] is a generic, scalable, and efficient multimodal pre-training approach - for vision-language pre-training - that can enable any family of LLMs to understand images while keeping their parameters frozen. It combines frozen pre-trained image models and language models, achieving state-of-the-art performance on various vision-language tasks, while having fewer trainable parameters than other existing methods. BLIP-2 stands out for its computing efficiency, thanks to its use of frozen unimodal models and a lightweight Q-Former (Querying Transformer), introduced to bridge the gap between vision and language modalities.

The BLIP-2 model is structured with three components: a frozen image encoder, a frozen Large Language Model (LLM), and a Querying Transformer (Q-Former). The frozen image encoder and LLM process visual and text inputs, respectively, while the Q-Former bridges the gap between them by extracting relevant visual features from the image encoder based on the text instruction. The pre-training strategy involves two stages: in the vision-and-language representation learning stage, the Q-Former aligns representations by learning to extract relevant image features for corresponding text; in the vision-to-language generative learning stage, it focuses on generating accurate text responses.

BLIP-2 uses the same pre-training dataset as BLIP [72] with 129M images in total, including COCO [23], VisualGenome [64], CC3M [112], CC12M [20], SBU [93], and 115M images from the LAION400M [109] dataset.

To conclude, BLIP-2 effectively combines pre-trained unimodal models and introduces a Querying Transformer to align visual and textual information, demonstrating a promising approach for vision-language tasks.

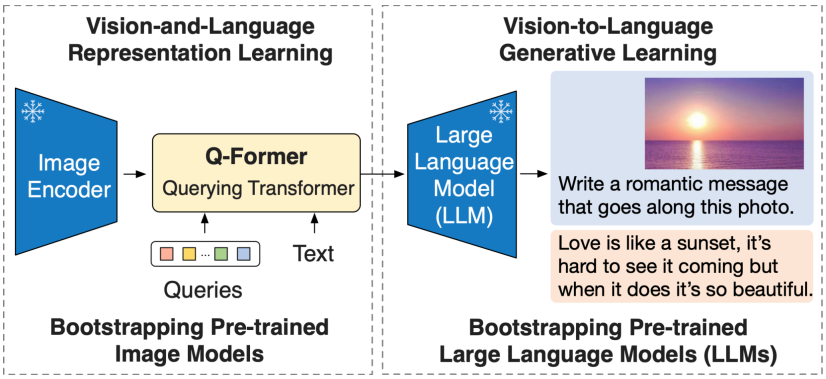


Figure 6: General Overview | Framework of BLIP-2 [71]

CLIP Contrastive Language-Image Pre-training [101], or CLIP, is a vision and language model developed by OpenAI. It is a multimodal and zero-shot model; the CLIP model utilizes a dual-encoder architecture to learn joint representations of images and text. As observed in Table 1, dual encoders processes visual and textual representations independently, and they use a simple mechanism, cosine similarity in this specific case, for the visual and textual module outputs to interact. With the technique of contrastive learning, CLIP is trained to understand that similar representations should be close to the latent space, while dissimilar ones should be far apart.

The CLIP model is a neural network model built on hundreds of millions of images and captions. In fact, the model has been jointly trained on the WebImageText dataset, a set of 400 million paired image-text pairs crawled form the web. CLIP can be adapted to perform a wide variety of tasks without needing additional training examples, while other models are not able to perform 'out of the box'. Its impressive zero-shot capabilities - zero-shot learning is the ability of the model to perform tasks that it was not explicitly programmed to do - make it able to accurately predict entire classes it has never seen before. In order for images and text to be connected to one another, they must both be embedded. To achieve this, the CLIP model consists of two sub-models called encoders: a text encoder that will embed text into mathematical space, and an image encoder that will embed images into mathematical space. Each image will go into the image encoder and the output for each image will also be a series of numbers;

the same for each text. One way to measure the 'goodness' of the model is how close the embedded representation - series of numbers - for each text is to the embedded representation for each image, and there is a convenient way to calculate the similarity between two series of numbers: the cosine similarity. Formally, for pre-training, CLIP is trained to identify the actual (image, text) pairings within a batch, chosen from the NxN possible combinations. The model learns a multi-modal embedding space by jointly training an image encoder and text encoder, aiming to maximize the cosine similarity between the embeddings of images and texts corresponding to the N authentic pairs in the batch, while simultaneously minimizing the cosine similarity between the embeddings of the $N^2 - N$ incorrect pairings. Ultimately, the optimization process involves minimizing a symmetric cross-entropy loss computed over these similarity scores [101].

CLIP was designed to mitigate a number of major problems in the standard deep learning approach to computer vision [101]. Deep learning, in fact, needs a lot of data, and vision models have traditionally been trained on manually labeled datasets that are expensive to construct and only provide supervision for a limited number of predetermined visual concepts, while CLIP learns from text-image pairs that are already publicly available on the internet. In addition, there is a gap between benchmark performance and real performance: this gap occurs because the models 'cheat' by only optimizing for performance on the benchmark, while CLIP can be evaluated on benchmarks without having to train on their data, so it can't 'cheat' in this manner.

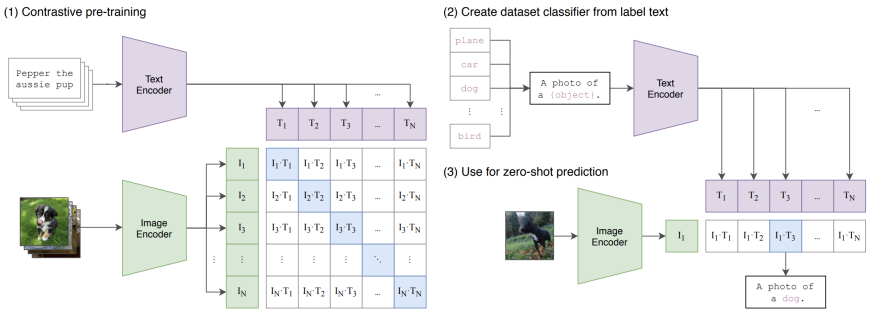


Figure 7: General Overview | Framework of CLIP [101]

For both the classification and retrieval tasks, the same models and versions of the models were used. Specifically, through the LAVIS library [67] [68], the relevant features were extracted for each model and then utilized in the processes of the two tasks. For CLIP, the model type utilized was ViT-B-32; for ALBEF, the pre-trained 'base' model was considered; for BLIP-2, the pre-trained 'pretrain' model was used.

3.3 Research Questions and Methods

Two sets of experiments were conducted to evaluate and measure bias in response to two research questions, involving the examination of the previously introduced models (ALBEF, BLIP-2, and CLIP). The following sections provide detailed explanations for each experimental setting, but before delving into these settings, some concepts that will be useful later on will be defined.

Zero-shot learning is a paradigm in machine learning where a model is trained to recognize and perform tasks on classes or concepts it has never seen during the training phase [135]. Unlike traditional machine learning approaches that require labeled examples for each class, zero-shot learning aims to extend the model's capabilities to new, unseen classes. So, in the context of VL models, a zero-shot setting refers to a scenario where the model performs certain tasks, such as classification or retrieval in the case of this work, for classes or concepts that were not part of its training data. In a zero-shot setting, the model relies on its ability to understand semantic relationships between different classes or concepts, and this often involves leveraging auxiliary information, such as textual descriptions or embeddings, to make predictions on classes not encountered during the training phase.

In the context of bias and vision-language models, evaluating their performances on the zero-shot setting enables consideration of biases acquired during training. Zero-shot learning facilitates assessment of the model's capacity to generalize to unseen classes or concepts, encompassing both the dataset and categories under study, as well as the target concepts. This helps in understanding how well the model can adapt to new

tasks or concepts, which is important for real-world applications in situations where the complete range of potential categories or concepts may not be known in advance. In this case, the zero-shot setting can be particularly useful due to the absence of labeled data and the novelty of the dataset to be used, which has obviously not been utilized in the training of the models. Furthermore, through the evaluation of performance on zero-shot tasks, insights into the model's comprehension of semantic relationships between classes or concepts can be obtained.

When evaluating or considering tasks involving more than one modality, it is important to clarify the distinction between multimodal and cross-modal. While both terms involve multiple modalities, multimodal generally implies the joint consideration of information from different modalities, whereas cross-modal specifically highlights the interaction or transfer of information across modalities.

As stated in Garrido-Muñoz et al. [37], a machine learning system may be deemed unbiased or fair when its predictions do not favor members of any relevant population group or discriminate against any other. A machine learning system is assessed as fair if and only if the scores it assigns to different subgroups (gender and race in this case) do not differ substantially. More formally, in a bias study, the subgroups under study, also known as target entities, may be represented as sets of instances $X = x_1, x_2, ..., x_N$ and $Y = y_1, y_2, ..., y_N$. X may be images, and the attributes towards which the bias is being measured may be given as sets $A = a_1, a_2, ..., a_M$ and $B = b_1, b_2, ..., b_M$ [59]. A machine learning model is then said to be fair towards subgroups X and Y with respect to attributes A and B if and only if $\phi(X, A, B) \approx \phi(Y, A, B)$, where ϕ is some scoring function that scores the similarity of the sets of attributes A, B to a target entity X or Y.

Zero-Shot Retrieval The primary objective of the majority of cross-modal algorithms is to find a latent space representation for the two modalities such that the relevant data pairs come closer to each other in this space [32]. The main challenge, in fact, is known as the heterogeneity gap and it arises because items from different modalities have distinct data types, making direct measurement of similarity impossible. Consequently,

most methods aim to bridge this gap by learning a latent representation space where the similarity between items from different modalities can be measured. In the zero-shot setting, cross-modal retrieval aims to retrieve relevant instances or data points from a dataset without direct exposure to examples of the target class during training. Zero-Shot Retrieval is the task of finding relevant items across different modalities without having received any training examples [48]; the model is expected to understand the similarities and relationships between different classes to perform effective retrieval for previously unseen classes [69]. For instance, given an image, find relevant text and vice versa.

A pertinent research question for this task can be summarized as follows: *"To what extent does race and gender bias exist in the zero-shot retrieval task involving text and image modalities in vision-language models? Are specific demographic groups consistently misrepresented or underrepresented in this scenario?"*

The challenge of generating fair and unbiased image retrieval results given neutral textual queries - with no explicit gender or race connotations - is addressed in this task. Utilizing the constructed morphed version of the UTKFace dataset as a starting point, the evaluation of zero-shot retrieval involves constructing a benchmark with twelve gender-race neutral queries with the following structure: *"A photo of a {bias word} person"*, whereas the bias word is taken from the presented list of bias words. The intuitive idea is that each group, both for race and gender, comprises the same individuals, making the results independent of these factors. The vision-language models under examination are ALBEF, BLIP-2, and CLIP, and the concept of equal opportunity is evaluated, obtaining $Bias@k$ scores. The fairness measure "Equal Opportunity" was introduced by Hardt et al. [45] in 2016, and Wang et al. [128] adapted its definition to the task of image retrieval. In particular, a fair retrieval result is that given the T_B, the retrieval text, the probability of occurrence of a sample with a certain attribute in R_k, the retrieval result, is close to the probability distribution of this sample group in the whole dataset. This definition can be summarised in the following formula: $Bias@k(A, T_B) = |p_{A,T_B,R_k} - p_{A,T,R}|$; where p_{A,T_B,R_k} represents the occurrence

probability of samples with attribute A in R_k when T_B is used as the retrieval text, and $p_{A,T,R}$ represents the occurrence probability of samples with attribute A in the whole image dataset. Another metric will be utilized: the Wasserstein distance. The Wasserstein distance, also known as the earth mover's distance (EMD), is a measure of the dissimilarity between two probability distributions. Specifically, it quantifies the minimum amount of work required to transform one distribution into another, where 'work' is typically interpreted as the amount of mass that must be moved, with the cost of moving mass proportional to the distance it is moved. In the context of this task, this metric serves as a powerful tool for quantifying the differences between probability distributions, which represent the counts of categories for each prompt. For instance, in evaluating race, we'll analyze the top k retrieved images for each prompt and assess the distribution of race categories in the retrieval process. Basically, we'll evaluate how the distribution from each model differs from a benchmark distribution.

Zero-shot Classification Classification based on zero-shot learning involves the ability of a model to assign inputs into novel classes on which the model has not previously seen any training examples [104]. Zero-Shot Classification is in fact a challenging task that requires the model to categorize inputs using information from both images and text in a zero-shot setting. This demands the model to leverage its learned knowledge to generalize and make accurate predictions for entirely new classes based on semantic relationships acquired during training [27] [59].

A pertinent research question for this task can be summarized as follows: *"Is there evidence of race and gender bias in zero-shot classification tasks performed by vision-language models, and if so, how does this bias manifest across different demographic groups? Are specific demographic groups consistently misclassified in this setting?"*

Starting from the constructed morphed version of the UTKFace dataset, this task aims to explore potential biases in the classification process, even when each considered group is composed of the same individuals altered according to race and gender, by utilizing bias words and analyze if models potentially relate specific groups to specific terms. The methodology involves performing zero-shot classification of target images to

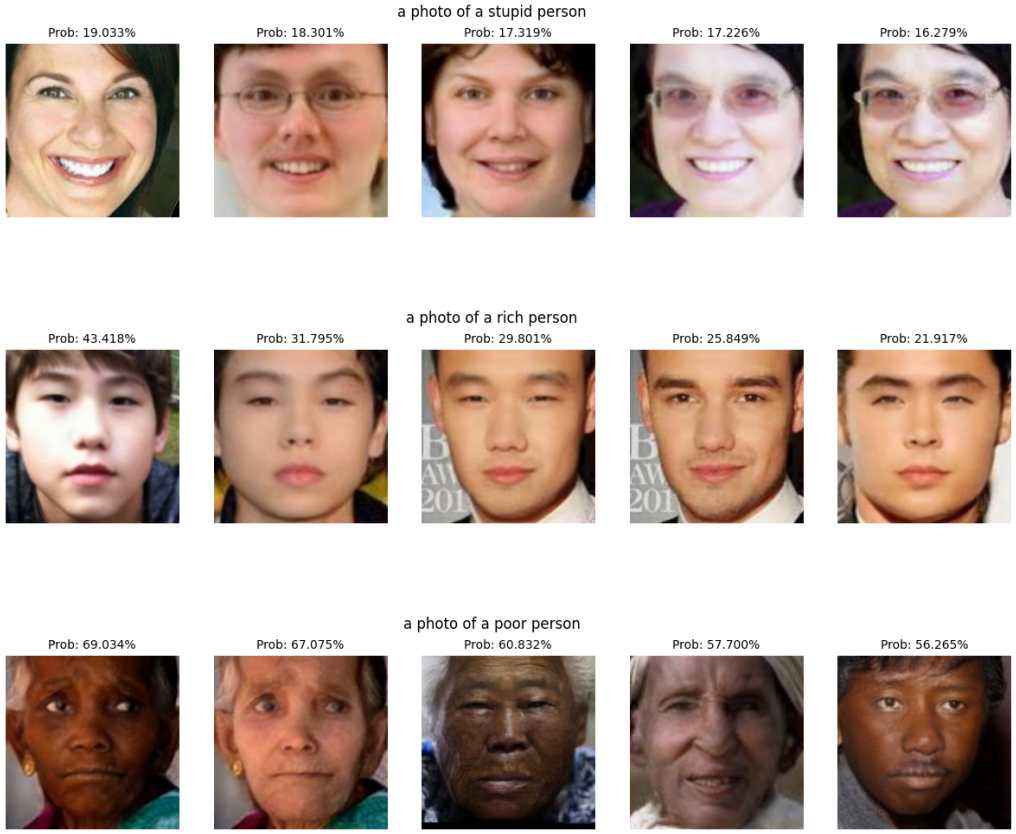
attribute words. For each image group, the average probabilities of classified words are returned. Even in this case, the vision-language models under examination are ALBEF, BLIP-2, and CLIP, and the output will be composed by the top closest attributes (most associated words) for each (race and gender) category, with respective probabilities.

4 Experimental Evaluations

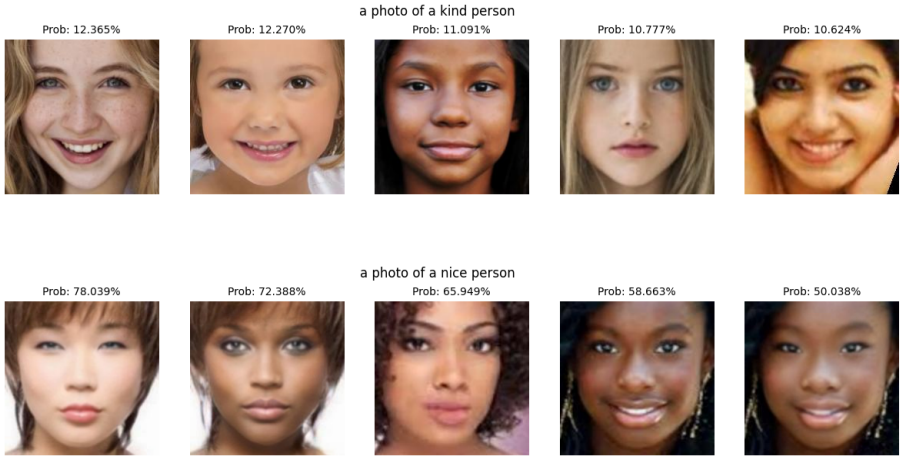
Before delving into the heart of the experiments and analyzing the results obtained, let's briefly recap what will be used and the intuitive idea behind these experiments. The idea is that, for each race and gender, the same individuals are depicted, and the models should be able to assess the individuals, not the race-related - or gender-related - characteristics they observe in the images. Considering the previously constructed morphed dataset, where each person is altered to reflect the other races or gender, the tasks of zero-shot classification and zero-shot retrieval aim to identify the potential presence of bias, assessing and quantifying how vision and language models perform these tasks. In the case of classification, by observing how the models associate image categories with different bias words; in the case of retrieval, by observing the distribution of extracted categories for each prompt. Both experiments will be conducted using the two morphed datasets for race and gender, allowing for the sequential assessment of race followed by gender.

4.1 Zero-Shot Retrieval

Since the categories under study contain the same people but with altered attributes, we do not expect to observe substantial and significant differences in the retrieval task, thereby not observing biases related to race and gender. This task was performed with three values of k, namely $k = 4$, $k = 20$ and $k = 50$, to evaluate the eventual presence of bias in the retrieval task at different levels.



In these examples above, the model's capability - in this case CLIP and in the race scenario - emerges not only to rely solely on characteristics such as skin color to evaluate the probability of extraction with the prompt, but rather to extract two or three times the same morphed image. However, bias also emerges in associating only women, for example, when extracting 'A photo of a stupid person,' or only men when extracting 'A photo of a rich person,' where predominantly Asian men - both original and morphed - are extracted. The presence of bias exists in both gender and race, sometimes simultaneously and not as one maybe could expect, especially if we consider that the tasks are conducted separately for gender and race categories. Examples abound for all models and mainly all prompts. For instance, ALBEF retrieves only images of women when prompted to find 'A photo of a kind person' in the gender scenario, and BLIP-2 do the same for 'A photo of a nice person'. This tells us that these models also pick up biases such as that empathy-related adjectives are connected with females.



The results are presented in the tables below, with some examples of the most interesting count tables highlighted in the styled tables at the end of this paragraph.

$k = 4$	*Race*			*Gender*		
	Mean AD	Median AD	WD	Mean AD	Median AD	WD
ALBEF	0.87	1	0.22	3	2	0.37
BLIP-2	0.54	0.5	0.13	2.17	1	0.27
CLIP	0.92	1	0.23	2.5	1	0.31

Table 5: Equal Opportunity | Mean and Median absolute difference and Wasserstein distance of race and gender at *Bias*@4 of different VL pre-trained models

$k = 20$	*Race*			*Gender*		
	Mean AD	Median AD	WD	Mean AD	Median AD	WD
ALBEF	2.87	2.5	0.14	8.66	4	0.22
BLIP-2	2.17	2	0.11	9.17	3.5	0.23
CLIP	2.79	2.25	0.14	7.17	3	0.18

Table 6: Equal Opportunity | Mean and Median absolute difference and Wasserstein distance of race and gender at *Bias*@20 of different VL pre-trained models

$k = 50$	*Race*			*Gender*		
	Mean AD	Median AD	WD	Mean AD	Median AD	WD
ALBEF	5.81	5.25	0.12	17.2	8	0.17
BLIP-2	4.06	3.75	0.08	20.5	8	0.20
CLIP	5.21	3.75	0.10	13.3	4.5	0.13

Table 7: Equal Opportunity | Mean and Median absolute difference and Wasserstein distance of race and gender at *Bias*@50 of different VL pre-trained models

Table 5, 6 and 7 reports the $Bias@k$scores for the selected vision-language models, where the first column refers to the mean absolute difference and the second to the median absolute difference, while the third column refers to the Wasserstein distance. Here, the already explained concept of equal opportunity is central, as in theory - for fair results - the proportion of images retrieved should follow the proportion of images in the dataset. After pre-processing, image features are matched with text features using similarity computations and softmax normalization to obtain probabilities, the results are stored in a dictionary associating each prompt with a list of image paths and probabilities. The dataframe is ordered according to the probability value, and the first k images are extracted for further analysis. For instance, in the case of race and $k = 20$, the metrics are calculated as follows. The average absolute difference is calculated by summing the absolute differences between the counts of each race category and $k/4$ (indicating the expected count), across all prompts, and then dividing by the total number of prompts; the median is calculated by finding the median of the absolute differences between the counts of each race category and across all prompts. A similar approach is used in the case of gender. In this way, we obtain values - shown in the tables above - that lets us compare the presence of bias in the selected VL models. In addition, the Wasserstein distance lets us compare the actual distribution of race/gender category counts obtained from the retrieval task to the distribution we would ideally expect if the retrieval process was perfectly balanced. This metric serves as a quantitative measure of the dissimilarity between these two distributions - our actual distribution and an ideal distribution -, helping to assess the fairness of the retrieval process for each VL model. When comparing models using the Wasserstein distance, it's relevant to do so within the same value of k. This is because a lower k value amplifies even minor distribution changes, resulting in greater deviations from the ideal distribution. Conversely, when k is higher, this effect is less pronounced.

A curious trend that emerges from the results observable in the three tables is certainly the ability of the BLIP-2 model to produce fewer biases when it comes to race, but conversely, to be more prone to biases when it comes to gender. Regarding race bias,

in fact, the best results are obtained with BLIP-2, while for gender bias the best results are achieved with CLIP. ALBEF, however, is the model that achieves the worst performances related to bias in both scenarios. Notably, these results highlight the significant bias exhibited by BLIP-2 in zero-shot retrieval tasks, particularly concerning gender bias; this bias becomes more pronounced with increasing values of k. Additionally, the general performance of ALBEF warrants attention, especially in relation to both race and gender biases.

Table 8: ALBEF | Race | k20

Prompts	**Asian Count**	**Black Count**	**Indian Count**	**White Count**
a photo of a criminal person	2	7	9	2
a photo of a dishonest person	2	10	3	5
a photo of a evil person	1	3	3	13
a photo of a kind person	1	9	6	4
a photo of a lawful person	7	1	1	11
a photo of a nice person	2	5	1	12
a photo of a poor person	6	9	4	1
a photo of a rich person	6	4	5	5
a photo of a smart person	4	2	5	9
a photo of a stupid person	3	0	0	17
a photo of a terrible person	5	3	3	9
a photo of a trustworthy person	3	4	4	9

Table 9: BLIP-2 | Gender | k20

Prompts	**Female Count**	**Male Count**
a photo of a criminal person	8	12
a photo of a dishonest person	7	13
a photo of an evil person	19	1
a photo of a kind person	19	1
a photo of a lawful person	12	8
a photo of a nice person	11	9
a photo of a poor person	11	9
a photo of a rich person	4	16
a photo of a smart person	2	18
a photo of a stupid person	6	14
a photo of a terrible person	8	12
a photo of a trustworthy person	18	2

Table 10: BLIP-2 | Gender | k50

Prompts	**Female Count**	**Male Count**
a photo of a criminal person	18	32
a photo of a dishonest person	23	27
a photo of an evil person	44	6
a photo of a kind person	44	6
a photo of a lawful person	28	22
a photo of a nice person	29	21
a photo of a poor person	28	22
a photo of a rich person	13	37
a photo of a smart person	3	47
a photo of a stupid person	18	32
a photo of a terrible person	16	34
a photo of a trustworthy person	41	9

Table 11: CLIP | Race | k50

Prompts	**Asian Count**	**Black Count**	**Indian Count**	**White Count**
a photo of a criminal person	9	14	16	11
a photo of a dishonest person	14	13	9	14
a photo of a evil person	10	7	17	16
a photo of a kind person	15	14	16	5
a photo of a lawful person	8	21	10	11
a photo of a nice person	19	10	11	10
a photo of a poor person	8	15	17	10
a photo of a rich person	33	5	5	7
a photo of a smart person	18	10	16	6
a photo of a stupid person	11	5	9	25
a photo of a terrible person	6	3	1	40
a photo of a trustworthy person	8	12	22	8

In the presented tables, cells highlighted in dark yellow represent values significantly higher than those in the corresponding columns of the same row, while blue cells indicate lower values.

Table 8 displays the distribution of race categories across various prompts, for $k = 20$ and the ALBEF model. Notably, the White category consistently exhibits the highest counts, especially in prompts associated with terms like 'evil', 'lawful', 'nice', and 'stupid'. Conversely, the Black category predominates in instances related to 'dishon-

est'.

Table 9 and Table 10 present the distribution of gender categories across various prompts generated by the BLIP-2 model, corresponding to $k = 20$ and $k = 50$ respectively. Despite the difference in the number of retrieved images (k), the overall trends remain consistent between the two tables, with similar patterns evident in the highlighted cells. A notable trend observed in both tables is the association of the male category with the term 'smart'. Conversely, the female category tends to predominate in prompts related to 'evil', 'kind', and 'trustworthy'. These consistent associations suggest certain biases or societal stereotypes embedded in the model's outputs. Additionally, it's interesting to note the model's lack of awareness regarding the connections and relationships between adjectives. In this case, for instance, the results of BLIP-2 denote the association of bias words like 'kind' and 'trustworthy' predominantly with the female category, but the same category being associated with 'evil' contradicts these assumptions.

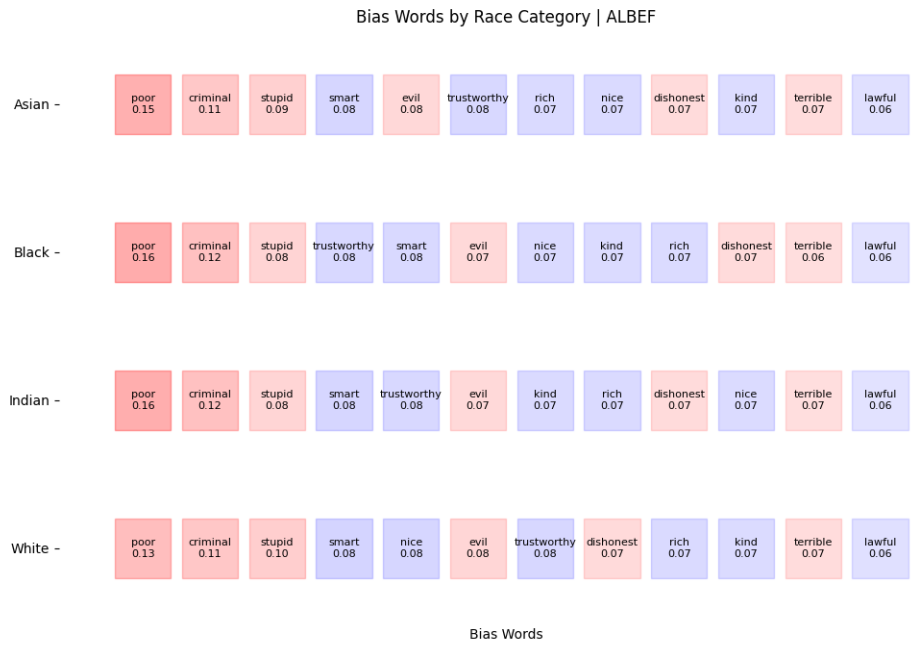
As displayed in Table 11, the results obtained from CLIP demonstrate a lower presence of bias, with only a few noteworthy observations. In particular, the Asian category exhibits a notably high count for the 'rich' prompt, while the White category displays high counts for 'terrible' and low counts for 'kind' prompts. These findings suggest a relatively balanced representation across racial categories compared to other models.

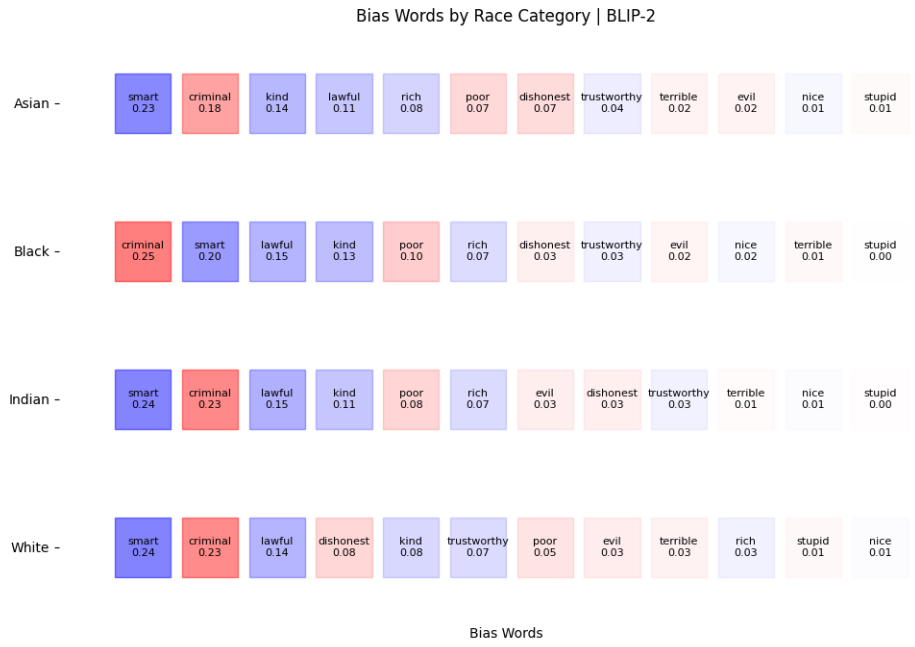
In summary, the analysis of multiple models across various prompts consistently reveals discernible trends and patterns regarding biases and stereotypes, particularly concerning gender.

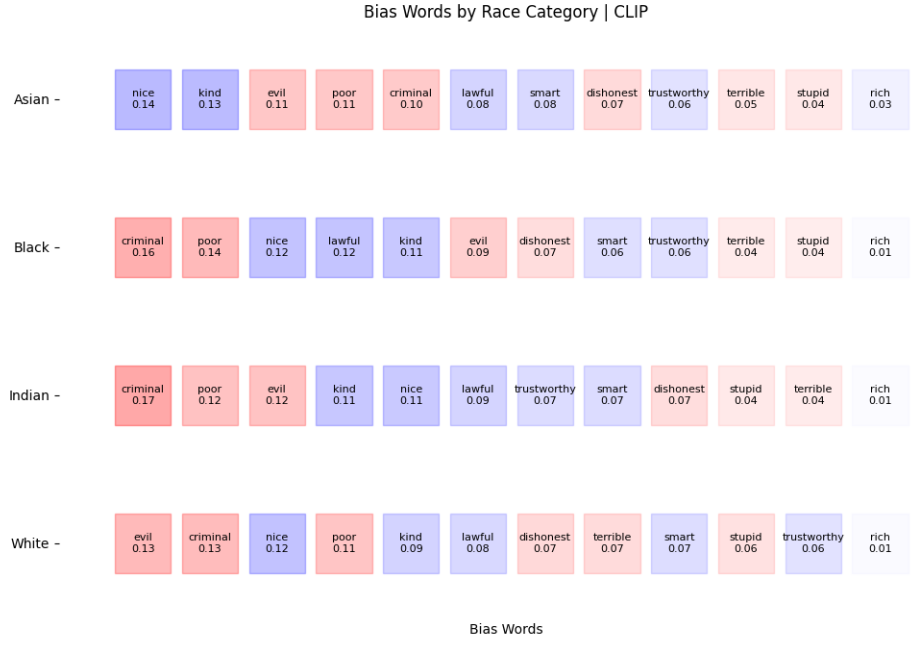
4.2 Zero-Shot Classification

The zero-shot classification task was performed to investigate the association of images and bias words, with the aim of exploiting the presence of bias for race and gender. For the three models and for each category, average probabilities were obtained with which each category was associated with each bias word. Even in this case, since the

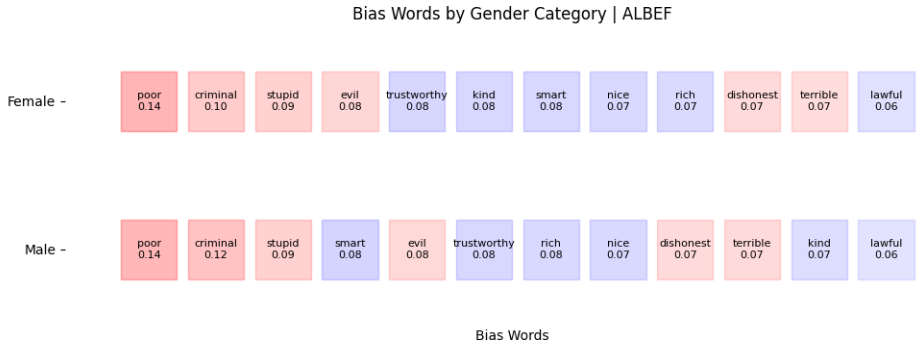
categories under study contain the same people but with altered attributes, we expect to not observe relevant different magnitudes of associations between categories and bias words. The process involves iterating over different - race and gender - categories initially, where images from each category are processed along with the predefined bias words. Features are then extracted for each of the three selected VL models, followed by the calculation of similarity scores between image and text features, and then probabilities for predefined bias labels are computed based on these scores. Finally, mean probabilities for each label across all images in each category are calculated. The results are illustrated in the visualizations, with the first block dedicated to race and the subsequent block focusing on gender. In both cases, bias words are distinguished by color; positive associations are represented in blue, while negative associations are depicted in red. Additionally, the hue of each color corresponds to the intensity and value of the association, providing a clear representation of the relationships observed.

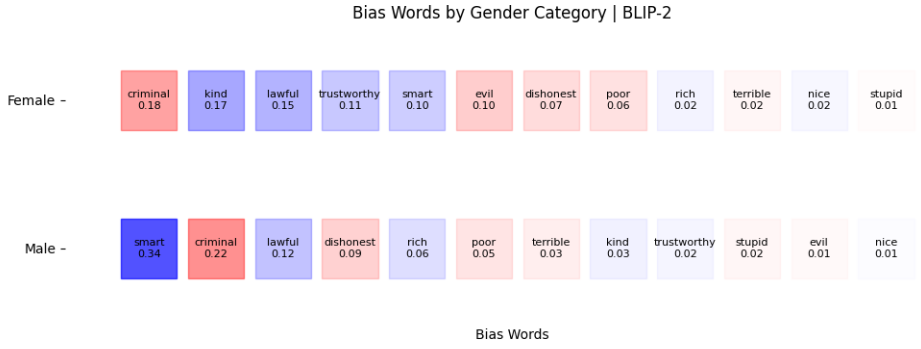


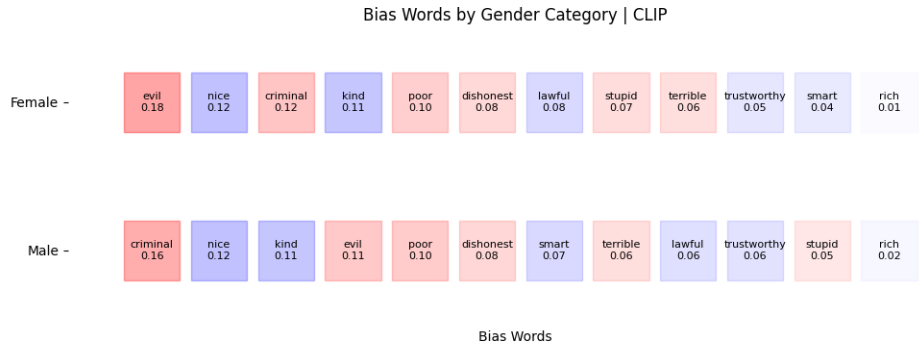

Bias Words by Race Category | BLIP-2
Asian
smart 0.23
criminal 0.18
kind 0.14
lawful 0.11
rich 0.08
poor 0.07
dishonest 0.07
trustworthy 0.04
terrible 0.02
evil 0.02
nice 0.01
stupid 0.01
Black
criminal 0.25
smart 0.20
lawful 0.15
kind 0.13
poor 0.10
rich 0.07
dishonest 0.03
trustworthy 0.03
evil 0.02
nice 0.02
terrible 0.01
stupid 0.00
Indian
smart 0.24
criminal 0.23
lawful 0.15
kind 0.11
poor 0.08
rich 0.07
evil 0.03
dishonest 0.03
trustworthy 0.03
terrible 0.01
nice 0.01
stupid 0.00
White
smart 0.24
criminal 0.23
lawful 0.14
dishonest 0.08
kind 0.08
trustworthy 0.07
poor 0.05
evil 0.03
terrible 0.03
rich 0.03
stupid 0.01
nice 0.01
Bias Words


Bias Words by Race Category | CLIP
Asian
nice 0.14
kind 0.13
evil 0.11
poor 0.11
criminal 0.10
lawful 0.08
smart 0.08
dishonest 0.07
trustworthy 0.06
terrible 0.05
stupid 0.04
rich 0.03
Black
criminal 0.16
poor 0.14
nice 0.12
lawful 0.12
kind 0.11
evil 0.09
dishonest 0.07
smart 0.06
trustworthy 0.06
terrible 0.04
stupid 0.04
rich 0.01
Indian
criminal 0.17
poor 0.12
evil 0.12
kind 0.11
nice 0.11
lawful 0.09
trustworthy 0.07
smart 0.07
dishonest 0.07
stupid 0.04
terrible 0.04
rich 0.01
White
evil 0.13
criminal 0.13
nice 0.12
poor 0.11
kind 0.09
lawful 0.08
dishonest 0.07
terrible 0.07
smart 0.07
stupid 0.06
trustworthy 0.06
rich 0.01
Bias Words

Across all three selected VL models, there is a consistent trend of higher associations towards negative attributes. In general, positive bias words received lower probabilities, indicating a systemic imbalance in the models' classifications. Another thing that stands out is the difference in intensity of the results produced by the BLIP-2 model compared to the other two models, for both race and gender. In BLIP-2, indeed, there is a greater imbalance among the various words, it exhibits a pronounced difference in association with specific terms within each category, strongly associating with certain terms while rarely associating with others. Across all pre-trained models, some bias words - like 'criminal' - have high values for all the race categories.







In the selected vision-language models, there are more noticeable differences in gender bias compared to the race bias observed above. The term 'smart', for instance, shows higher associations with males across all three models, and while it is more unbalanced in BLIP-2, it may indicate the presence of a gender bias associating intelligence more strongly with males. Additionally, attributes like 'evil' and 'criminal' exhibit differences in associations between genders across models. Even in the case of gender, the ALBEF model predominantly associates negative words with both categories, but in general the differences between genders are not pronounced. Here as well, the BLIP-2 model exhibits strong associations with certain words and sparse associations with others. Even the CLIP model shows differences in the values associated with common stereotypes, although not very significant.

In summary, the zero-shot classification task offers some interesting insights into the classification tendencies of the selected vision-language models, as it reveals some differences between the association values of bias words across categories, especially in the case of gender.

5 Conclusions

This thesis aimed to investigate biases concerning race and gender in selected vision-language (VL) models. The methodology involved constructing a new dataset of images by altering race-related and gender-related attributes from the UTKFace dataset, resulting in new images of the same individuals but differing only in the chosen dimension. This morphed dataset was then utilized to investigate the presence of bias within three selected VL models, namely ALBEF, BLIP-2 and CLIP. Through the selection of specific bias words and the execution of tasks like zero-shot retrieval and zero-shot classification, the aim was to analyze, quantify, and explore the presence of the previously mentioned biases within the selected models. The underlying idea behind the two tasks was that, with the same individuals in the image dataset, no differences in the results of the various tasks were expected based on ethnicity or gender. The twelve selected bias words were used in both tasks, in the case of zero-shot retrieval to compose the race-neutral and gender-neutral queries with the following structure: *'A photo of a {} person'*; while in the case of zero-shot classification, target images were classified to attribute words obtaining the probability of association of each image with the twelve words. The goal in both cases was to assess any differences, first for race and then for gender, in the probabilities - of retrieval and classification - for certain terms and identify any patterns characterized by the presence of bias.

Considering the research questions that drove the primary focus of this work, it is possible to make some considerations and provide some answers, as the experimental evaluations conducted revealed some interesting insights into the behavior of the selected vision-language (VL) models.

This study discovered significant tendencies of VL models to associate certain attributes with specific demographics in both race and gender scenarios in the zero-shot retrieval task. Biases were evident in the distribution of extracted categories in zero-shot retrieval, indicating potential disparities in model performance based on race and gender. Specifically, the BLIP-2 model exhibited fewer biases in race but pronounced biases

in gender, while ALBEF showed poor performance in both scenarios. Despite expectations of uniform performance across different demographic attributes in the tasks, biases were evident in the distribution of extracted categories in zero-shot retrieval for the three values of k - the number of images retrieved - considered, indicating potential disparities in model performance based on race and gender.

Regarding evidence of race and gender bias in zero-shot classification tasks, the study found a consistent trend of higher associations towards negative attributes across all selected vision-language models. This indicates systemic biases in the classification tendencies of these models, with certain demographic groups consistently misclassified. The BLIP-2 model, in particular, exhibited a greater imbalance in associating certain terms within each category, suggesting a propensity for strong associations with specific attributes.

The observed patterns in the performance of the selected vision-language models can be attributed to several factors, although definitive conclusions require further investigations. Various factors may contribute to ALBEF's overall poorer performance. From its underlying architecture, which plays a significant role in determining its capabilities, to the objectives and loss functions employed during training, which directly influence its behavior, to the fine-tuning strategies utilized during training or adaptation to specific tasks, as well as the model's complexity in terms of parameters and layers, further impact its efficacy. Of particular importance is the quality, quantity, and diversity of the training data, which heavily influences model performance. Since ALBEF was trained on a smaller dataset compared to BLIP-2 and CLIP, it might not have captured the nuances of different demographic groups as effectively, thus allowing biases to persist more prominently.

The findings of this work align with previous research indicating persistent biases in VL models. While some models showed improvements in exhibiting biases, significant challenges remain in achieving fairness and equity. The conducted research contributes to the existing knowledge by highlighting the presence and the complexities of bias in vision-language models and their implications for fairness and inclusivity, and the

observed biases underscore the need for continued research and development of strategies to address biases in artificial intelligence systems effectively. Understanding these biases is essential for developing fairer AI systems and promoting equitable outcomes in various applications, considering that the use of these models is expected to increase in the coming years.

While the experimental evaluations provided some valuable insights, it's important to recognize the inherent limitations of this study and identify areas for potential improvement. This work focused on specific bias categories, leaving room for exploration for other demographic attributes and categories that may contribute to bias. A more comprehensive understanding of bias in vision-language models can be achieved by exploring more and alternative demographic attributes, such as age or socioeconomic status, or by including different categories apart from race and gender, such as sexual orientation, disabilities, and much more. Increasing the size of the dataset by adding more images could also strengthen the robustness of the results. Additionally, exploring different sets of bias words and incorporating additional tasks could offer deeper insights into the behavior of vision-language models. Finally, considering alternative approaches to investigate multimodal behavior could provide complementary perspectives on how these models process and interpret visual and textual information.

In conclusion, this study underscores the importance of addressing biases in vision-language models to ensure fairness and equity in AI applications. By understanding and mitigating biases, we can in fact advance the development of more inclusive and ethical AI systems. The research findings directly address the initial research questions outlined in the introduction, providing insights into biases in VL models and their implications for fairness and inclusivity. The findings of this research call for continued efforts in research, policy, and practice to promote fairness and accountability in AI technologies. Promoting awareness and accountability in AI development and deployment is in fact crucial for creating a more equitable digital future.

References

[1] Ehsan Abbasnejad et al. "Counterfactual Vision and Language Learning". In: *2020 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2020, pp. 10041–10051. DOI: 10.1109/CVPR42600.2020.01006.

[2] Sandhini Agarwal et al. *Evaluating CLIP: Towards Characterization of Broader Capabilities and Downstream Implications*. 2021. arXiv: 2108.02818.

[3] Aishwarya Agrawal et al. *VQA: Visual Question Answering*. 2016. arXiv: 1505.00468 [cs.CL].

[4] Sayak Paul Alara Dirik. *A Dive into Vision-Language Models*. 2023. URL: https://huggingface.co/blog/vision_language_pretraining (visited on 08/06/2023).

[5] Jean-Baptiste Alayrac et al. *Flamingo: a Visual Language Model for Few-Shot Learning*. 2022. arXiv: 2204.14198.

[6] Ricardo Baeza-Yates. "Bias on the web". In: *Communications of the ACM* 61 (May 2018), pp. 54–61. DOI: 10.1145/3209581.

[7] Hangbo Bao et al. *VLMo: Unified Vision-Language Pre-Training with Mixture-of-Modality-Experts*. 2022. arXiv: 2111.02358.

[8] Hugo Elias Berg et al. "A Prompt Array Keeps the Bias Away: Debiasing Vision-Language Models with Adversarial Learning". In: *AACL*. 2022.

[9] Shruti Bhargava and David Forsyth. *Exposing and Correcting the Gender Bias in Image Captioning Datasets and Models*. 2019. arXiv: 1912.00578.

[10] Federico Bianchi et al. "Easily Accessible Text-to-Image Generation Amplifies Demographic Stereotypes at Large Scale". In: *2023 ACM Conference on Fairness, Accountability, and Transparency*. ACM, 2023. DOI: 10.1145/3593013.3594095. URL: https://doi.org/10.1145%2F3593013.3594095.

[11] Yonatan Bitton et al. "WinoGAViL: Gamified association benchmark to challenge vision-and-language models". In: *Advances in Neural Information Processing Systems* 35 (2022), pp. 26549–26564.

[12] Tolga Bolukbasi et al. "Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings". In: *Advances in Neural Information Processing Systems*. Ed. by D. Lee et al. Vol. 29. Curran Associates, Inc., 2016. URL: https://proceedings.neurips.cc/paper_files/paper/2016/file/a486cd07e4ac3d270571622f4f316ec5-Paper.pdf.

[13] Shikha Bordia and Samuel R. Bowman. "Identifying and Reducing Gender Bias in Word-Level Language Models". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*. Minneapolis, Minnesota: Association for Computational Linguistics, 2019, pp. 7–15. DOI: 10.18653/v1/N19-3002. URL: https://aclanthology.org/N19-3002.

[14] Tom B. Brown et al. *Language Models are Few-Shot Learners*. 2020. arXiv: 2005.14165.

[15] Emanuele Bugliarello et al. *IGLUE: A Benchmark for Transfer Learning across Modalities, Tasks, and Languages*. 2022. arXiv: 2201.11732 [cs.CL].

[16] Emanuele Bugliarello et al. *Measuring Progress in Fine-grained Vision-and-Language Understanding*. 2023. arXiv: 2305.07558 [cs.CL].

[17] Emanuele Bugliarello et al. *Multimodal Pretraining Unmasked: A Meta-Analysis and a Unified Framework of Vision-and-Language BERTs*. 2021. arXiv: 2011.15124 [cs.CL].

[18] Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. "Semantics derived automatically from language corpora contain human-like biases". In: *Science* 356.6334 (2017), pp. 183–186. DOI: 10.1126/science.aal4230. eprint: https://www.science.org/doi/pdf/10.1126/science.aal4230. URL: https://www.science.org/doi/abs/10.1126/science.aal4230.

[19] Feiqi Cao et al. *Understanding Attention for Vision-and-Language Tasks*. 2022. arXiv: 2208.08104.

[20] Soravit Changpinyo et al. *Conceptual 12M: Pushing Web-Scale Image-Text Pre-Training To Recognize Long-Tail Visual Concepts*. 2021. arXiv: 2102.08981.

[21] FL. Chen, DZ. Zhang, and ML. et al Han. "VLP: A Survey on Vision-language Pre-training". In: vol. 20. 2023, pp. 38–56. DOI: 10.1007/s11633-022-1369-5. URL: https://link.springer.com/article/10.1007/s11633-022-1369-5.

[22] Xi Chen et al. *PaLI: A Jointly-Scaled Multilingual Language-Image Model*. 2023. arXiv: 2209.06794.

[23] Xinlei Chen et al. *Microsoft COCO Captions: Data Collection and Evaluation Server*. 2015. arXiv: 1504.00325.

[24] Yen-Chun Chen et al. *UNITER: UNiversal Image-TExt Representation Learning*. 2020. arXiv: 1909.11740.

[25] Jaemin Cho et al. *Unifying Vision-and-Language Tasks via Text Generation*. 2021. arXiv: 2102.02779 [cs.CL].

[26] Ching-Yao Chuang et al. "Debiasing Vision-Language Models via Biased Prompts". In: *ArXiv* abs/2302.00070 (2023).

[27] Nassim Dehouche. "Implicit Stereotypes in Pre-Trained Classifiers". In: *IEEE Access* 9 (2021), pp. 167936–167947. DOI: 10.1109/ACCESS.2021.3136898.

[28] Karan Desai et al. "RedCaps: Web-curated image-text data created by the people, for the people". In: *NeurIPS Datasets and Benchmarks*. 2021.

[29] Jacob Devlin et al. *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. 2019. arXiv: 1810.04805 [cs.CL].

[30] Shizhe Diao et al. *Write and Paint: Generative Vision-Language Models are Unified Modal Learners*. 2023. arXiv: 2206.07699.

[31] Yifan Du et al. *A Survey of Vision-Language Pre-Trained Models*. 2022. arXiv: 2202.10936.

[32] Titir Dutta and Soma Biswas. "Generalized Zero-Shot Cross-Modal Retrieval". In: *IEEE Transactions on Image Processing* PP (June 2019), pp. 1–1. DOI: 10.1109/TIP.2019.2923287.

[33] Clayton Fields and Casey Kennington. *Vision Language Transformers: A Survey*. 2023. arXiv: 2307.03254.

[34] Zhe Gan et al. *Vision-Language Pre-training: Basics, Recent Advances, and Future Trends*. 2022. arXiv: 2210.09263.

[35] Noa Garcia et al. "Uncurated Image-Text Datasets: Shedding Light on Demographic Bias". In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2023, pp. 6957–6966.

[36] Nikhil Garg et al. "Word embeddings quantify 100 years of gender and ethnic stereotypes". In: *Proceedings of the National Academy of Sciences* 115.16 (2018). DOI: 10.1073/pnas.1720347115. URL: https://doi.org/10.1073%2Fpnas.1720347115.

[37] Ismael Garrido-Muñoz et al. "A Survey on Bias in Deep NLP". In: *Applied Sciences* 11.7 (2021). ISSN: 2076-3417. DOI: 10.3390/app11073184. URL: https://www.mdpi.com/2076-3417/11/7/3184.

[38] Jindong Gu et al. *A Systematic Survey of Prompt Engineering on Vision-Language Foundation Models*. 2023. arXiv: 2307.12980.

[39] Jing Gu et al. "Vision-and-Language Navigation: A Survey of Tasks, Methods, and Future Directions". In: *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 2022. DOI: 10.18653/v1/2022.acl-long.524. URL: http://dx.doi.org/10.18653/v1/2022.acl-long.524.

[40] Meng-Hao Guo et al. "Attention mechanisms in computer vision: A survey". In: *Computational Visual Media* 8 (2021), pp. 331–368. URL: https://api.semanticscholar.org/CorpusID:244117862.

[41] Meng-Hao Guo et al. "Attention mechanisms in computer vision: A survey". In: *Computational Visual Media* (2022), pp. 1–38.

[42] Wei Guo and Aylin Caliskan. "Detecting Emergent Intersectional Biases: Contextualized Word Embeddings Contain a Distribution of Human-like Biases". In: *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*. AIES '21. Virtual Event, USA: Association for Computing Machinery, 2021, pp. 122–133. ISBN: 9781450384735. DOI: 10.1145/3461702.3462536. URL: https://doi.org/10.1145/3461702.3462536.

[43] Tanmay Gupta et al. *Towards General Purpose Vision Systems*. 2022. arXiv: 2104.00743.

[44] Melissa Hall et al. "Vision-Language Models Performing Zero-Shot Tasks Exhibit Gender-based Disparities". In: *ArXiv* abs/2301.11100 (2023).

[45] Moritz Hardt, Eric Price, and Nathan Srebro. *Equality of Opportunity in Supervised Learning*. 2016. arXiv: 1610.02413 [cs.LG].

[46] Sophia Harrison, Eleonora Gualdoni, and Gemma Boleda. *Run Like a Girl! Sports-Related Gender Bias in Language and Vision*. 2023. arXiv: 2305.14468.

[47] Kaiming He et al. *Momentum Contrast for Unsupervised Visual Representation Learning*. 2020. arXiv: 1911.05722.

[48] Mariya Hendriksen et al. *Scene-centric vs. Object-centric Image-Text Cross-modal Retrieval: A Reproducibility Study*. 2023. arXiv: 2301.05174 [cs.IR].

[49] Yusuke Hirota, Yuta Nakashima, and Noa Garcia. "Gender and Racial Bias in Visual Question Answering Datasets". In: *2022 ACM Conference on Fairness, Accountability, and Transparency*. ACM, 2022. DOI: 10.1145/3531146.3533184. URL: https://doi.org/10.1145%2F3531146.3533184.

[50] Xin Hong et al. *Transformation Driven Visual Reasoning*. 2021. arXiv: 2011.13160 [cs.CV].

[51] Dirk Hovy, Federico Bianchi, and Tommaso Fornaciari. ""You Sound Just Like Your Father" Commercial Machine Translation Systems Include Stylistic Biases". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Ed. by Dan Jurafsky et al. Online: Association for Computational Linguistics, 2020, pp. 1686–1690. DOI: 10.18653/v1/2020.acl-main.154. URL: https://aclanthology.org/2020.acl-main.154.

[52] Xiaowei Hu et al. *Scaling Up Vision-Language Pre-training for Image Captioning*. 2022. arXiv: 2111.12233.

[53] Zhicheng Huang et al. *Pixel-BERT: Aligning Image Pixels with Text by Deep Multi-Modal Transformers*. 2020. arXiv: 2004.00849.

[54] Zhicheng Huang et al. *Seeing Out of tHe bOx: End-to-End Pre-training for Vision-Language Representation Learning*. 2021. arXiv: 2104.03135.

[55] Yuqi Huo et al. *WenLan: Bridging Vision and Language by Large-Scale Multi-Modal Pre-Training*. 2021. arXiv: 2103.06561.

[56] Ben Hutchinson et al. "Social Biases in NLP Models as Barriers for Persons with Disabilities". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 2020, pp. 5491–5501. DOI: 10.18653/v1/2020.acl-main.487. URL: https://aclanthology.org/2020.acl-main.487.

[57] Phillip Isola et al. *Image-to-Image Translation with Conditional Adversarial Networks*. 2018. arXiv: 1611.07004.

[58] Jiho Jang et al. *Unifying Vision-Language Representation Space with Single-tower Transformer*. 2022. arXiv: 2211.11153 [cs.LG].

[59] Sepehr Janghorbani and Gerard de Melo. *Multi-Modal Bias: Introducing a Framework for Stereotypical Bias Assessment beyond Gender and Race in Vision Language Models*. 2023. arXiv: 2303.12734.

[60] Chao Jia et al. *Scaling Up Visual and Vision-Language Representation Learning With Noisy Text Supervision*. 2021. arXiv: 2102.05918.

[61] Kimmo Karkkainen and Jungseock Joo. "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*. 2021, pp. 1548–1558.

[62] Wonjae Kim, Bokyung Son, and Ildoo Kim. *ViLT: Vision-and-Language Transformer Without Convolution or Region Supervision*. 2021. arXiv: 2102.03334 [stat.ML].

[63] Ronak Kosti et al. "Context Based Emotion Recognition using EMOTIC Dataset". In: *IEEE Transactions on Pattern Analysis and Machine Intelligence* (2019), pp. 1–1. DOI: 10.1109/tpami.2019.2916866. URL: https://doi.org/10.1109%2Ftpami.2019.2916866.

[64] Ranjay Krishna et al. *Visual Genome: Connecting Language and Vision Using Crowdsourced Dense Image Annotations*. 2016. arXiv: 1602.07332.

[65] Jiyoung Lee et al. *Context-Aware Emotion Recognition Networks*. 2019. arXiv: 1908.05913.

[66] Chenliang Li et al. *mPLUG: Effective and Efficient Vision-Language Learning by Cross-modal Skip-connections*. 2022. arXiv: 2205.12005 [cs.CL].

[67] Dongxu Li et al. *LAVIS: A Library for Language-Vision Intelligence*. 2022. arXiv: 2209.09019.

[68] Dongxu Li et al. "LAVIS: A One-stop Library for Language-Vision Intelligence". In: *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*. Toronto, Canada: Association for Computational Linguistics, July 2023, pp. 31–41. URL: https://aclanthology.org/2023.acl-demo.3.

[69] Fengling Li et al. *Cross-Modal Retrieval: A Systematic Review of Methods and Future Directions*. 2023. arXiv: 2308.14263 [cs.IR].

[70] Junnan Li et al. *Align before Fuse: Vision and Language Representation Learning with Momentum Distillation*. 2021. arXiv: 2107.07651.

[71] Junnan Li et al. *BLIP-2: Bootstrapping Language-Image Pre-training with Frozen Image Encoders and Large Language Models*. 2023. arXiv: 2301.12597.

[72] Junnan Li et al. *BLIP: Bootstrapping Language-Image Pre-training for Unified Vision-Language Understanding and Generation*. 2022. arXiv: 2201.12086.

[73] Liunian Harold Li et al. *VisualBERT: A Simple and Performant Baseline for Vision and Language*. 2019. arXiv: 1908.03557.

[74] Wei Li et al. *UNIMO: Towards Unified-Modal Understanding and Generation via Cross-Modal Contrastive Learning*. 2022. arXiv: 2012.15409 [cs.CL].

[75] Xiujun Li et al. *Oscar: Object-Semantics Aligned Pre-training for Vision-Language Tasks*. 2020. arXiv: 2004.06165.

[76] Yikai Li, C. L. Philip Chen, and Tong Zhang. "A Survey on Siamese Network: Methodologies, Applications, and Opportunities". In: *IEEE Transactions on Artificial Intelligence* 3.6 (2022), pp. 994–1014. DOI: 10.1109/TAI.2022.3207112.

[77] Zewen Li et al. *A Survey of Convolutional Neural Networks: Analysis, Applications, and Prospects*. 2020. arXiv: 2004.02806.

[78] Dingkang Liang et al. *CrowdCLIP: Unsupervised Crowd Counting via Vision-Language Model*. 2023. arXiv: 2304.04231.

[79] FaceApp Technology Limited. *FaceApp (Version 11.9.2) [Mobile App]*. 2023. URL: https://www.faceapp.com/.

[80] Fangyu Liu et al. *Visually Grounded Reasoning across Languages and Cultures*. 2021. arXiv: 2109.13238 [cs.CL].

[81] Yinhan Liu et al. *RoBERTa: A Robustly Optimized BERT Pretraining Approach*. 2019. arXiv: 1907.11692 [cs.CL].

[82] Jiasen Lu et al. *ViLBERT: Pretraining Task-Agnostic Visiolinguistic Representations for Vision-and-Language Tasks*. 2019. arXiv: 1908.02265.

[83] Zixian Ma et al. "CREPE: Can Vision-Language Foundation Models Reason Compositionally?" In: *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*. 2023, pp. 10910–10921.

[84] Thomas Manzini et al. "Black is to Criminal as Caucasian is to Police: Detecting and Removing Multiclass Bias in Word Embeddings". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, 2019, pp. 615–621. DOI: 10.18653/v1/N19-1062. URL: https://aclanthology.org/N19-1062.

[85] Junhua Mao. "Multimodal Learning with Vision and Language". In: *2019 Ninth International Conference on Image Processing Theory, Tools and Applications (IPTA)* (2019), pp. i–i. URL: https://api.semanticscholar.org/CorpusID:28017608.

[86] Ninareh Mehrabi et al. *A Survey on Bias and Fairness in Machine Learning*. 2022. arXiv: 1908.09635 [cs.LG].

[87] Emiel van Miltenburg, Desmond Elliott, and Piek Vossen. "Talking about other people: an endless range of possibilities". In: *Proceedings of the 11th International Conference on Natural Language Generation*. Tilburg University, The Netherlands: Association for Computational Linguistics, 2018, pp. 415–420. DOI: 10.18653/v1/W18-6550. URL: https://aclanthology.org/W18-6550.

[88] Aditya Mogadala, Marimuthu Kalimuthu, and Dietrich Klakow. "Trends in Integration of Vision and Language Research: A Survey of Tasks, Datasets, and Methods". In: *Journal of Artificial Intelligence Research* 71 (2021), pp. 1183–1317. DOI: 10.1613/jair.1.11688. URL: https://doi.org/10.1613%2Fjair.1.11688.

[89] Moin Nadeem, Anna Bethke, and Siva Reddy. *StereoSet: Measuring stereotypical bias in pretrained language models*. 2020. arXiv: 2004.09456 [cs.CL].

[90] Yulei Niu et al. *Counterfactual VQA: A Cause-Effect Look at Language Bias*. 2021. arXiv: 2006.04315.

[91] Alexandra Olteanu et al. "Social Data: Biases, Methodological Pitfalls, and Ethical Boundaries". In: *Frontiers in Big Data* 2 (2019). ISSN: 2624-909X. DOI: 10.3389/fdata.2019.00013. URL: https://www.frontiersin.org/articles/10.3389/fdata.2019.00013.

[92] OpenAI et al. *GPT-4 Technical Report*. 2023. arXiv: 2303.08774.

[93] Vicente Ordonez, Girish Kulkarni, and Tamara Berg. "Im2Text: Describing Images Using 1 Million Captioned Photographs". In: *Advances in Neural Information Processing Systems*. Ed. by J. Shawe-Taylor et al. Vol. 24. Curran Associates, Inc., 2011. URL: https://proceedings.neurips.cc/paper_files/paper/2011/file/5dd9db5e033da9c6fb5ba83c7a7ebea9-Paper.pdf.

[94] Felix Ott et al. "Auxiliary Cross-Modal Representation Learning With Triplet Loss Functions for Online Handwriting Recognition". In: *IEEE Access* 11 (2023), pp. 94148–94172. DOI: 10.1109/access.2023.3310819. URL: https://doi.org/10.1109%2Faccess.2023.3310819.

[95] Wilkins EG Pannucci CJ. "Identifying and avoiding bias in research". In: *"Plast Reconstr Surg."*. 2010, 126(2):619–625. DOI: 10.1097/PRS.0b013e3181de24bc.

[96] Letitia Parcalabescu et al. *VALSE: A Task-Independent Benchmark for Vision and Language Models Centered on Linguistic Phenomena*. 2022. arXiv: 2112.07566 [cs.CL].

[97] Kunyu Peng et al. *Affect-DML: Context-Aware One-Shot Recognition of Human Affect using Deep Metric Learning*. 2021. arXiv: 2111.15271.

[98] Ioannis Pikoulis, Panagiotis P. Filntisis, and Petros Maragos. "Leveraging Semantic Scene Characteristics and Multi-Stream Convolutional Architectures in a Contextual Approach for Video-Based Visual Emotion Recognition in the Wild". In: *2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021)*. IEEE, 2021. DOI: 10.1109/fg52635.2021.9666957. URL: https://doi.org/10.1109%2Ffg52635.2021.9666957.

[99] Jordi Pont-Tuset et al. "Connecting Vision and Language with Localized Narratives". In: *ECCV*. 2020.

[100] Alec Radford et al. "Improving language understanding by generative pre-training". In: (2018).

[101] Alec Radford et al. *Learning Transferable Visual Models From Natural Language Supervision*. 2021. arXiv: 2103.00020.

[102] Scott Reed et al. *Generative Adversarial Text to Image Synthesis*. 2016. arXiv: 1605.05396 [cs.NE].

[103] Candace Ross, Boris Katz, and Andrei Barbu. "Measuring Social Biases in Grounded Vision and Language Embeddings". In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Online: Association for Computational Linguistics, 2021, pp. 998–1008. DOI: 10.18653/v1/2021.naacl-main.78. URL: https://aclanthology.org/2021.naacl-main.78.

[104] Samuele Ruffino et al. *Zero-shot Classification using Hyperdimensional Computing*. 2024. arXiv: 2401.16876.

[105] Gabriele Ruggeri and Debora Nozza. "A Multi-dimensional study on Bias in Vision-Language models". In: *Findings of the Association for Computational Linguistics: ACL 2023*. Toronto, Canada: Association for Computational Linguistics, 2023, pp. 6445–6455. URL: https://aclanthology.org/2023.findings-acl.403.

[106] Nripsuta Saxena et al. *How Do Fairness Definitions Fare? Examining Public Attitudes Towards Algorithmic Definitions of Fairness*. 2019. arXiv: 1811.03654 [cs.AI].

[107] Madeline Chantry Schiappa et al. *Probing Conceptual Understanding of Large Visual-Language Models*. 2023. arXiv: 2304.03659.

[108] Robin M. Schmidt. *Recurrent Neural Networks (RNNs): A gentle Introduction and Overview*. 2019. arXiv: 1912.05911 [cs.LG].

[109] Christoph Schuhmann et al. "LAION-400M: Open Dataset of CLIP-Filtered 400 Million Image-Text Pairs". In: *CoRR* abs/2111.02114 (2021). arXiv: 2111.02114. URL: https://arxiv.org/abs/2111.02114.

[110] Christoph Schuhmann et al. *LAION-5B: An open large-scale dataset for training next generation image-text models*. 2022. arXiv: 2210.08402.

[111] Ashish Seth, Mayur Hemani, and Chirag Agarwal. "DeAR: Debiasing Vision-Language Models with Additive Residuals". In: *ArXiv* abs/2303.10431 (2023).

[112] Piyush Sharma et al. "Conceptual Captions: A Cleaned, Hypernymed, Image Alt-text Dataset For Automatic Image Captioning". In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics, 2018, pp. 2556–2565. DOI: 10.18653/v1/P18-1238. URL: https://aclanthology.org/P18-1238.

[113] Ravi Shekhar et al. "FOIL it! Find One mismatch between Image and Language caption". In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, 2017. DOI: 10.18653/v1/p17-1024. URL: https://arxiv.org/abs/1705.01359.

[114] Amanpreet Singh et al. *FLAVA: A Foundational Language And Vision Alignment Model*. 2022. arXiv: 2112.04482.

[115] Brandon Smith et al. "Balancing the Picture: Debiasing Vision-Language Datasets with Synthetic Contrast Sets". In: *ArXiv* abs/2305.15407 (2023).

[116] Krishna Srinivasan et al. "WIT: Wikipedia-based Image Text Dataset for Multimodal Multilingual Machine Learning". In: *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 2021. DOI: 10.1145/3404835.3463257. URL: https://doi.org/10.1145%2F3404835.3463257.

[117] Tejas Srinivasan and Yonatan Bisk. "Worst of Both Worlds: Biases Compound in Pre-trained Vision-and-Language Models". In: *ArXiv* abs/2104.08666 (2021).

[118] Weijie Su et al. *VL-BERT: Pre-training of Generic Visual-Linguistic Representations*. 2020. arXiv: 1908.08530.

[119] Harini Suresh and John Guttag. "A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle". In: *Equity and Access in Algorithms, Mechanisms, and Optimization*. ACM, 2021. DOI: 10.1145/3465416.3483305. URL: https://doi.org/10.1145%2F3465416.3483305.

[120] Hao Tan and Mohit Bansal. *LXMERT: Learning Cross-Modality Encoder Representations from Transformers*. 2019. arXiv: 1908.07490 [cs.CL].

[121] Bart Thomee et al. "YFCC100M". In: *Communications of the ACM* 59.2 (2016), pp. 64–73. DOI: 10.1145/2812802. URL: https://doi.org/10.1145%2F2812802.

[122] Tristan Thrush et al. *Winoground: Probing Vision and Language Models for Visio-Linguistic Compositionality*. 2022. arXiv: 2204.03162.

[123] Shagun Uppal et al. *Multimodal Research in Vision and Language: A Review of Current and Emerging Trends*. 2020. arXiv: 2010.09522.

[124] Ashish Vaswani et al. *Attention Is All You Need*. 2023. arXiv: 1706.03762 [cs.CL].

[125] Jialu Wang, Yang Liu, and Xin Eric Wang. *Are Gender-Neutral Queries Really Gender-Neutral? Mitigating Gender Bias in Image Search*. 2021. arXiv: 2109.05433.

[126] Jialu Wang, Yang Liu, and Xin Eric Wang. *Assessing Multilingual Fairness in Pre-trained Multimodal Representations*. 2022. arXiv: 2106.06683 [cs.CL].

[127] Junke Wang et al. *OmniVL:One Foundation Model for Image-Language and Video-Language Tasks*. 2022. arXiv: 2209.07526.

[128] Junyang Wang, Yi Zhang, and Jitao Sang. *FairCLIP: Social Bias Elimination based on Attribute Prototype Learning and Representation Neutralization*. 2022. arXiv: 2210.14562.

[129] Peng Wang et al. *OFA: Unifying Architectures, Tasks, and Modalities Through a Simple Sequence-to-Sequence Learning Framework*. 2022. arXiv: 2202.03052.

[130] Wenhui Wang et al. *Image as a Foreign Language: BEiT Pretraining for All Vision and Vision-Language Tasks*. 2022. arXiv: 2208.10442.

[131] Zirui Wang et al. *SimVLM: Simple Visual Language Model Pretraining with Weak Supervision*. 2022. arXiv: 2108.10904.

[132] Laura Weidinger et al. *Ethical and social risks of harm from Language Models*. 2021. arXiv: 2112.04359 [cs.CL].

[133] Benjamin Wilson, Judy Hoffman, and Jamie Morgenstern. *Predictive Inequity in Object Detection*. 2019. arXiv: 1902.11097.

[134] Robert Wolfe et al. "Contrastive Language-Vision AI Models Pretrained on Web-Scraped Multimodal Data Exhibit Sexual Objectification Bias". In: *2023 ACM Conference on Fairness, Accountability, and Transparency*. ACM, 2023. DOI: 10.1145/3593013.3594072. URL: https://doi.org/10.1145%2F3593013.3594072.

[135] Yongqin Xian et al. *Zero-Shot Learning – A Comprehensive Evaluation of the Good, the Bad and the Ugly*. 2020. arXiv: 1707.00600.

[136] Sang Michael Xie et al. *An Explanation of In-context Learning as Implicit Bayesian Inference*. 2022. arXiv: 2111.02080 [cs.CL].

[137] Haiyang Xu et al. *E2E-VLP: End-to-End Vision-Language Pre-training Enhanced by Visual Learning*. 2021. arXiv: 2106.01804.

[138] Kaiyuan Xu et al. *Project Implicit Demo Website Datasets*. 2024. DOI: 10.17605/OSF.IO/Y9HIQ. URL: osf.io/y9hiq.

[139] Kelvin Xu et al. "Show, Attend and Tell: Neural Image Caption Generation with Visual Attention". In: *CoRR* abs/1502.03044 (2015). arXiv: 1502.03044. URL: http://arxiv.org/abs/1502.03044.

[140] Xiao Xu et al. *BridgeTower: Building Bridges Between Encoders in Vision-Language Representation Learning*. 2023. arXiv: 2206.08657.

[141] Jianwei Yang et al. *Unified Contrastive Learning in Image-Text-Label Space*. 2022. arXiv: 2204.03610.

[142] Zhengyuan Yang et al. *UniTAB: Unifying Text and Box Outputs for Grounded Vision-Language Modeling*. 2022. arXiv: 2111.12085.

[143] Shaowei Yao and Xiaojun Wan. "Multimodal Transformer for Multimodal Machine Translation". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics, 2020, pp. 4346–4350. DOI: 10.18653/v1/2020.acl-main.400. URL: https://aclanthology.org/2020.acl-main.400.

[144] Gokul Yenduri et al. *Generative Pre-trained Transformer: A Comprehensive Review on Enabling Technologies, Potential Applications, Emerging Challenges, and Future Directions*. 2023. arXiv: 2305.10435.

[145] Jiahui Yu et al. *CoCa: Contrastive Captioners are Image-Text Foundation Models*. 2022. arXiv: 2205.01917.

[146] Desen Yuan. *Language bias in Visual Question Answering: A Survey and Taxonomy*. 2021. arXiv: 2111.08531.

[147] Lu Yuan et al. *Florence: A New Foundation Model for Computer Vision*. 2021. arXiv: 2111.11432.

[148] Mert Yuksekgonul et al. "When and Why Vision-Language Models Behave like Bags-Of-Words, and What to Do About It?" In: *The Eleventh International Conference on Learning Representations*. 2023. URL: https://openreview.net/forum?id=KRLUvxh8uaX.

[149] Rowan Zellers et al. *From Recognition to Cognition: Visual Commonsense Reasoning*. 2019. arXiv: 1811.10830 [cs.CV].

[150] Yan Zeng et al. *X^2-VLM: All-In-One Pre-trained Model For Vision-Language Tasks*. 2023. arXiv: 2211.12402.

[151] Zhixiong Zeng and Wenji Mao. *A Comprehensive Empirical Study of Vision-Language Pre-trained Model for Supervised Cross-Modal Retrieval*. 2022. arXiv: 2201.02772.

[152] Jingyi Zhang et al. *Vision-Language Models for Vision Tasks: A Survey*. 2023. arXiv: 2304.00685.

[153] Jingyi Zhang et al. *Vision-Language Models for Vision Tasks: A Survey*. 2024. arXiv: 2304.00685.

[154] Pengchuan Zhang et al. *VinVL: Revisiting Visual Representations in Vision-Language Models*. 2021. arXiv: 2101.00529.

[155] Yi Zhang, Junyang Wang, and Jitao Sang. *Counterfactually Measuring and Eliminating Social Bias in Vision-Language Pre-training Models*. 2022. arXiv: 2207.01056.

[156] Zhifei Zhang, Yang Song, and Hairong Qi. "Age Progression/Regression by Conditional Adversarial Autoencoder". In: *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE. 2017.

[157] Jieyu Zhao et al. "Gender Bias in Contextualized Word Embeddings". In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*. Minneapolis, Minnesota: Association for Computational Linguistics, 2019, pp. 629–634. DOI: 10.18653/v1/N19-1064. URL: https://aclanthology.org/N19-1064.

[158] Jieyu Zhao et al. *Men Also Like Shopping: Reducing Gender Bias Amplification using Corpus-level Constraints*. 2017. arXiv: 1707.09457 [cs.AI].

[159] Zijia Zhao et al. "MAMO: Masked Multimodal Modeling for Fine-Grained Vision-Language Representation Learning". In: *ArXiv* abs/2210.04183 (2022). URL: https://api.semanticscholar.org/CorpusID:252780916.

[160] Kaiyang Zhou et al. "Learning to Prompt for Vision-Language Models". In: *International Journal of Computer Vision* 130.9 (2022), pp. 2337–2348. DOI: 10.1007/s11263-022-01653-1. URL: https://doi.org/10.1007%2Fs11263-022-01653-1.

[161] Kankan Zhou, Eason Lai, and Jing Jiang. "VLStereoSet: A Study of Stereotypical Bias in Pre-trained Vision-Language Models". In: *AACL*. 2022.

[162] Wangchunshu Zhou et al. *VLUE: A Multi-Task Benchmark for Evaluating Vision-Language Models*. 2022. arXiv: 2205.15237.