

Exploring Vision Language Models for Facial Attribute Recognition: Emotion, Race, Gender, and Age

Nouar AlDahoul^a, Myles Joshua Toledo Tan^b, Harishwar Reddy Kasireddy^b, Yasir Zaki^{a,*}

^a*Computer Science Department, New York University Abu Dhabi Abu Dhabi, UAE*

^b*Department of Electrical and Computer Engineering, Herbert Wertheim College of Engineering University of Florida Florida, USA*

Abstract

Technologies for recognizing facial attributes like race, gender, age, and emotion have several applications, such as enhanced surveillance, personalized advertising content, sentiment analysis, interactive media, and the study of demographic trends and social behaviors. Analyzing demographic characteristics based on images and analyzing facial expressions have several challenges due to the complexity of humans' facial attributes and diversity in representation. Traditional approaches have employed convolutional neural networks (CNNs) and various other deep learning techniques, trained on extensive collections of labeled images. While these methods have demonstrated effective performance, there remains potential for further enhancements to increase the recognition accuracy. In this paper, we propose to utilize vision language models (VLMs) such as generative pre-trained transformer (GPT), Google GEMINI, large language and vision assistant (LLAVA), Google PaliGemma (Pathways Language and Image Model + Gemma model), and Microsoft Florence2 to recognize facial attributes such as race, gender, age group, and emotion from images with human faces. Various datasets like FairFace, AffectNet, and UTKFace have been utilized to evaluate the proposed solutions. The results show that VLMs are competitive and sometimes superior to traditional techniques. Additionally, we propose "FaceScanPaliGemma"—a fine-

*Corresponding author.

Email addresses: naa9497@nyu.edu (Nouar AlDahoul), tan.m@ufl.edu (Myles Joshua Toledo Tan), harishwarreddy.k@ufl.edu (Harishwar Reddy Kasireddy), yasir.zaki@nyu.edu (Yasir Zaki)

tuned PaliGemma model—for race, gender, age, and emotion recognition tasks. The results show an accuracy of 81.1% , 95.8%, 80%, and 59.4% for race, gender, age group, and emotion classification, respectively, outperforming pre-trained version of PaliGemma, other VLMs, and state-of-the-art methods. Finally, we propose “FaceScanGPT”, which is a GPT-4o model to recognize race, gender, age group, and emotion when several individuals are present in the image using a prompt engineered for a person with specific facial and/or physical attributes. The results underscore the superior multitasking capability of FaceScanGPT to detect the individual’s physical attributes like hair cut, clothing color, actions performed, postures, etc., and facial attributes like race, gender, etc., using only a prompt to drive the detection and recognition tasks.

Keywords: Vision Language Models, Facial Attribute Recognition, Generative Pre-trained Transformer, Pathways Language and Image

1. Introduction

Technologies for recognizing attributes such as race, gender, age, and emotion have various applications, including improved surveillance and personalized advertising content. Face attribute recognition is also utilized in demographic surveys conducted in marketing or social science research, with the goal of understanding human social behaviors and their connections to individuals’ demographic backgrounds [1]. Accurately identifying multiple attributes in an image presents several challenges due to the complexity of human attributes and diversity in representation. Human attributes like race, gender, age, and emotion are not always well-defined and can be subjective [2]. Machine learning (ML) models heavily rely on training data, and often biased or unrepresentative datasets can lead to biased or inaccurate predictions [3]. Attributes like age and emotion exhibit significant intra-class variability, meaning that different individuals within the same category (e.g., the same age group or emotion class) can have vastly different visual appearances [4].

In the area of image processing, artificial intelligence (AI), particularly ML, has made remarkable strides, transforming how computers understand and analyze visual data. Multi-task learning allows a single model to tackle multiple related tasks simultaneously, leveraging shared knowledge to improve performance across the board [5]. In image processing, multi-task

learning becomes indispensable, especially in the space of multi-attribute classification. Generative models represent a revolutionary step forward in artificial intelligence, particularly in the areas of natural language processing, vision, and generation. These models are designed to understand and generate human-like content, ushering in a new era of AI capabilities and applications.

Maintaining separate models for each individual task results in increased computational overhead and model complexity. Single-task classifiers are optimized to excel at a specific task but may struggle to generalize in unseen data or adapt to new tasks without retraining [6]. Many tasks in image processing share common underlying patterns and features. Single-task classifiers fail to leverage this shared information effectively, leading to suboptimal performance and redundant learning [7]. For instance, both facial recognition and facial attribute analysis tasks benefit from learning low-level visual features like edges and textures.

The critical role of ethics in AI development cannot be overstated, especially when it comes to applications involving sensitive human attributes such as race, gender, age, and emotion [8, 9]. As AI technologies become increasingly integrated into various aspects of society, it is imperative to prioritize ethical considerations to ensure fairness, transparency, and accountability. Many datasets used today to train AI models suffer from various biases, reflecting historical inequalities and societal prejudices [10, 11]. Even when trained on unbiased datasets, AI algorithms can perpetuate or amplify existing biases [12]. This can occur due to factors such as feature selection, model architecture, or optimization objectives.

The aim behind this work is to develop a multi-task classifier that leverages the capabilities of vision language models such as GPT [13], GEMINI [14], LLaVA [15], PaliGemma [16], and Florence-2 [17] generative models to simultaneously identify multiple human facial attributes—specifically race, gender, age, and emotion—from given images. This classifier integrates the advanced generative and understanding capabilities of these models to accurately and efficiently process and classify complex human's facial characteristics across diverse datasets. The creation of this multi-task classifier takes into account incorporating ethical AI practices, concentrating on reducing biases that new vision language models have recently addressed, ensuring privacy and consent, maintaining transparency and accountability, and promoting diversity and inclusion.

The use of VLM generative models significantly enhances the learning

capabilities of the proposed multi-task classifier. These models bring several key advantages:

1. VLMs such as GPT [13], GEMINI [14], LLaVA [15], PaliGemma [16], and Florence-2 [17] are designed to capture deep, nuanced representations of data. VLM, with its sophisticated language understanding capabilities, excels at interpreting and generating both textual and image-based information [18].
2. The pre-trained nature of these models utilizes transfer learning approach to offer a robust foundation of learned features that can be fine-tuned to specific tasks, such as the simultaneous classification of race, gender, age, and emotion [19].
3. VLM's architecture is particularly adept at understanding context, a critical factor when dealing with attributes that may be subtly expressed or highly context-dependent [20].
4. The models are scalable and flexible, allowing for continuous updates and adaptations as new data becomes available or as requirements evolve. This adaptability is crucial for keeping the classifier relevant and effective across different populations and evolving societal norms [21].

The approach of using vision language models for the multi-task classification of human attributes like race, gender, age, and emotion offers several significant benefits that address both the efficiency and effectiveness of AI systems:

1. Employing multi-task learning with vision language models leads to classifying multiple attributes simultaneously, rather than requiring separate models for each attribute, which helps to reduce computational resources [22].
2. VLM's advanced capabilities in handling multi-modal data and their profound contextual understanding enable them to adapt to the real-world diversity found in human attributes [23].
3. In traditional setups where separate models are trained for each task, there is often a significant overlap in what each model learns. VLM plays a crucial role in reducing this redundant learning [13].

In this study, our proposed facial attribute recognition solutions utilize state-of-the-art vision language models to recognize race, gender, age group, and emotion of persons using their face images. Our contributions can be summarized as follows:

- We formulated a facial attribute recognition task as a visual question answering task using various VLMs.
- We utilized public image datasets such as FairFace, AffectNet, and UTKFace, that have face images for various races, genders, age groups, and emotions for evaluation and comparison.
- We explored and evaluated the zero-shot classification, which is a significant capability existing in vision language models such as Google GEMINI 1.5, GPT-4o, LLaVA-NEXT, PaliGemma, and Florence-2, and employed it in the task of human’s facial attribute recognition.
- We fine-tuned PaliGemma utilizing FairFace and AffectNet datasets to improve the recognition accuracy. The outcome is FaceScanPaliGemma VLM.
- We explored the multitasking capability of GPT-4o using images with several persons having various physical and facial attributes.

This rest of the paper is organized as follows: In Section 2, we review previous works on facial attribute recognition methods. Section 3 presents our research motivation. In Section 4, we describe the datasets used to run the experiments. Section 5 discusses the experimental results and compares the proposed solution with other baseline methods. Finally, conclusions and future works are discussed in Section 6.

2. Related Work

2.1. Challenges in facial attribute datasets

The advancements in gender, race, age, and emotion classification applications requires datasets of diverse facial images that can address challenges such as imbalanced samples, pose variations, and varying lighting conditions. Recently, CNNs, especially the FaceNet model [24], have shown robustness in handling unbalanced data distributions [25]. In previous works, they identified a significant bias in public face image datasets, which predominantly feature Caucasian faces while considerably under-representing other racial groups, such as Latinx [1]. To address this issue of racial imbalance, they compiled a dataset, namely FairFace, consisting of 108,501 face images that is

balanced across different races [1]. They used this dataset to train a ResNet-34 [26] model to evaluate the classification performance for gender, race, and age. Similarly, recent works [27] used the same dataset with different models to improve the accuracy of the classification. They utilized VGGFace ResNet-50 convolutional neural network (VGGFace ResNet-50 CNN) [28] to extract the embedding vector from the face images in the FairFace dataset [27]. They added support vector machine (SVM) classifier after removing the top layers. Additionally, they explored numerous models including FaceNet+SVM [24], tuned EfficientNet-B7 [29], and large vision transformer [30] for race, and gender classification [27]. Furthermore, CLIP’s zero-shot classifier demonstrated their performance for race and gender classification using the FairFace dataset [31].

Existing annotated databases of facial expressions in the wild are small and include clean and high-quality posed facial expressions. However, posed expressions may not accurately represent many of the unopposed facial expressions encountered in daily life. As a result, a dataset, namely AffectNet [32], was proposed to be the largest database of facial expressions with annotations for eight expressions. Several works in literature have utilized the AffectNet dataset to train and evaluate their models for emotion or facial expression recognition [33, 34, 35, 36].

2.2. Deep Learning for facial attributes recognition

Race classification has become significant in applications like surveillance [37] and market advertising [38]. Recently, various deep learning models have been used for race identification [37, 39, 40, 41, 42].

Similarly, for gender recognition, deep learning techniques have improved the gender classification accuracy [43, 44, 45, 46, 47, 48]. Based on these techniques, the effectiveness of CNNs and AlexNet [49] under challenging conditions has been demonstrated, showing potential for applications of tracking and identification [45].

Additionally, age estimation from facial images is crucial for applications like security and social interaction. Several deep learning techniques have been proposed in the literature to improve feature extraction and thus enhance the accuracy of age classification [43, 46, 50, 51]. In this context, hybrid deep learning structures that combine CNNs and Extreme Learning Machine [52] have confirmed the efficacy of integrating multiple learning strategies [51].

Emotion recognition from facial expressions is crucial in human-computer interaction, with deep learning methods addressing challenges such as pose variations, illumination changes, and occlusions [53]. Based on this, transfer learning techniques using networks like ResNet50 [26] and VGG19 [54] have proven effective in combining feature extraction and classification for emotion recognition task [53].

Multimodal or multitask deep models have shown improvement in classification accuracy, demonstrating the utility of leveraging diverse biometric inputs [46, 47, 48]. The work aim to integrate multiple facial attributes extracted from facial images. This integration is crucial and has become a key area of interest in computer vision applications. CNN has been utilized to integrate gender and emotion [55]. Similarly, Visual attention-driven architectures have been employed for gender and ethnicity integration [56]. Additionally, the integration of age and gender prediction has enhanced biometric security and personalized systems.

Several studies have highlighted the efficiency of deep learning in enhancing real-time biometric recognition, paving the way for more compact, faster, and accurate systems. A multi-task CNN has been proposed for recognizing gender, age, ethnicity and emotion [57]. It improved efficiency in processing and memory usage while maintaining good accuracy across multiple tasks making it suitable for embedded systems [57] and for applications requiring real-time processing with limited computational resources [55].

Previously mentioned deep learning methods have outperformed traditional techniques in classifying race, gender, age, and emotion. However, there is still room for improvement. As such, this paper aims to improve facial attribute classification using VLMs.

2.3. Emerging role of LLMs in image processing

While LLMs are widely known for their expertise in language processing, recent studies have begun to explore their effectiveness in image recognition, often through multi-modal learning approaches [58]. Multimodal models like CLIP (contrastive language-image pretraining) [31] leverage the combination of LLMs and CNNs by learning from pairs of images and their textual descriptions. CLIP has achieved good performance on various image recognition benchmarks, including race, gender, and age recognition [31], thanks to its ability to create a shared representation space for both images and text, which enables robust performance even in zero-shot settings.

The use of large language models (LLMs), such as GPT [59] and BERT [60], has extended beyond text-based tasks, delving into areas like image recognition and processing. This broadening of application stems from the models' capacity to comprehend and generate human-like text, offering a fresh perspective on how images can be interpreted and analyzed [31].

Recent interdisciplinary research has started to investigate the potential of LLMs in image-related tasks such as construction processes, radiology, and medical visual question answering [61, 62, 63]. For instance, studies have shown that GPT can generate textual descriptions from images, paving the way for innovative approaches to image understanding through natural language [31].

LLMs have also been integrated with traditional vision models in tasks like visual question answering (VQA) [64], where models are trained to respond to questions based on image content. This highlights the synergy between LLMs and image recognition, requiring a deep understanding of both visual and textual data [64].

Integrating Large Language Models (LLMs) such as Google GEMINI 1.5 [65, 66], GPT-4o [59, 67], LLAVA-NEXT [68, 69, 70], PaliGemma [16], and Florence-2 [17] into facial recognition tasks can significantly enhance the performance and capabilities of deep learning models. This paper investigates the use of vision language models (VLMs) in addressing challenging computer vision tasks, such as emotion, gender, race, and age group classification. It leverages VLMs' capabilities in understanding and generation to improve visual recognition.

3. Research Motivation

The direct applications of vision language models (VLMs) in facial attribute recognition are still unexplored. The adaptability and contextual comprehension of VLMs hold promise for tackling complex challenges in facial attribute recognition, such as identifying race, gender, age, and emotions in face images that are blurred, noisy, have varying illumination, or have different facial orientations.

The existing solutions of VLMs [61, 62, 63, 71] in various applications indicate, with their sophisticated understanding and generative abilities, their potential capability to serve as complementary—or even alternative—solutions to traditional and CNN-based methods in facial attribute recognition systems. By harnessing the advanced language processing and context-

tual analysis capabilities of VLMs, researchers have the potential to achieve significant improvements in the accuracy, efficiency, and adaptability of technologies used for recognizing race, gender, age, and emotion.

4. Materials and Methods

This section describes the datasets used in the experiments conducted to evaluate the VLMs performance on facial attribute recognition tasks. Moreover, the section discusses the baseline methods usually used in the literature and highlights our proposed solutions.

4.1. Dataset Overview

4.1.1. FairFace dataset

To address the issue of racial bias in existing datasets, a face image dataset was created comprising of 108,501 images that is racially balanced [1]. The images were categorized into seven racial groups: White, Black, Indian, East Asian, Southeast Asian, Middle Eastern, and Latinx. These images were sourced from the YFCC-100M Flickr dataset [72] and annotated according to race, gender, and age groups. The dataset has binary gender classification: male and female, and several age groups: 0-2, 3-9, 10-19, 20-29, 30-39, 40-49, 50-59, 60-69, and 70+. The images have a resolution of 224 x 224 pixels. A few samples of the FairFace dataset are shown in Figure 1. We combined the age groups of the FairFace dataset into five categories that reflect different social and economic roles: 0-9, 10-19, 20-39, 40-59, 60+ and found that the dataset is imbalanced in terms of race and age groups that we selected. The numbers of samples for race, gender, and age group for the training and testing sets for each category are presented in Tables Tables 1 to 3. In this study, the FairFace dataset was employed to evaluate VLMs and, specifically, to assess the classification capabilities of FaceScanPaliGemma on race, gender, and age group.

4.1.2. Emotion dataset

AffectNet is a very challenging and extensive facial expression dataset containing approximately 0.4 million images that have been manually labeled to represent eight different facial expressions: neutral, happy, angry, sad, fear, surprise, disgust, and contempt [32]. The images were gathered from the Internet by conducting searches across three major search engines using 1,250 emotion-related keywords in six different languages. The images have a



Figure 1: Several samples from each race category to show the challenging contents of this dataset. The races are as follows: first row: Black, second row: Indian, third row: Middle Eastern, fourth row: Latinx_Hispanic, fifth row: White, sixth row: East Asian, and seventh row: Southeast Asian.

Race	Train samples	Test samples
Black	12,233	1556
East Asian	12,287	1550
Indian	12,319	1516
Latinx_Hispanic	13,367	1623
Middle_East	9216	1209
Southeast Asian	10,795	1415
White	16,527	2085
Total	86,744	10,954

Table 1: Number of samples in training and testing set in FairFace dataset per race category.

Gender	Train samples	Test samples
Male	45,986	5792
Female	40,758	5162
Total	86,744	10954

Table 2: number of samples in training/testing set in FairFace dataset per gender category

Age Group	Train samples	Test samples
0-9	12,200	1555
10-19	9103	1181
20-39	44,848	5630
40-59	16,972	2149
60+	3621	439
Total	86,744	10954

Table 3: Number of samples in training and testing set in FairFace dataset per age category.

resolution of 224 x 224 pixels. A few samples of AffectNet dataset are shown in Figure 2. AffectNet dataset is again imbalanced in terms of emotion or facial expression. The numbers of samples for the training and testing sets for each emotion category are shown presented in Table 4. In this study, the AffectNet dataset was employed to evaluate Vision-Language Models (VLMs) in general, with a specific focus on assessing FaceScanPaliGemma for emotion classification.

Emotion	Train samples	Test samples
neutral	74,874	500
happy	134,415	500
sad	25,459	500
surprise	14,090	500
fear	6,378	500
disgust	3,803	500
anger	24,882	500
contempt	3,750	499
Total	287,651	3999

Table 4: Number of samples in training/testing set in AffectNet dataset per emotion.

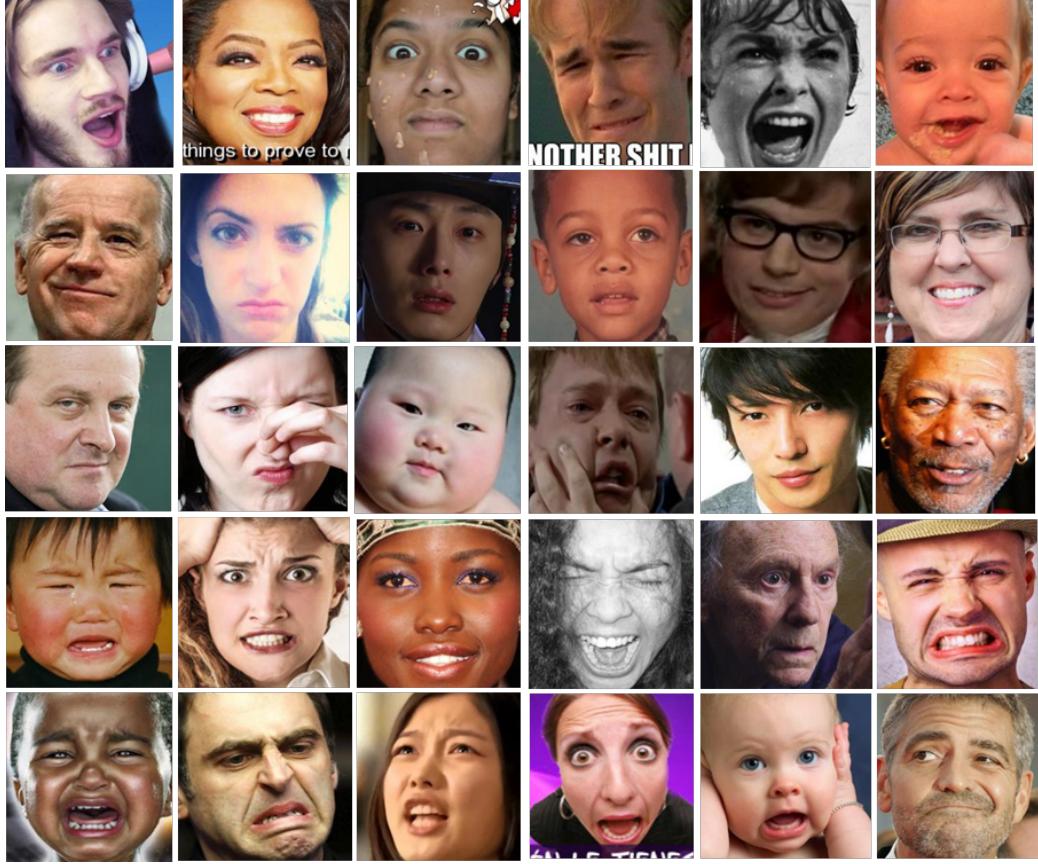


Figure 2: Several samples from the AffectNet dataset.

4.1.3. UTK-Face dataset

UTKFace dataset is a large-scale dataset consisting of over 20,000 images with annotations of age, gender, and ethnicity [73]. The images exhibit a wide range of variations in pose, facial expressions, lighting, occlusion, and resolution. This dataset is suitable for a variety of tasks, such as face detection, ethnicity classification, gender classification, age group classification, age estimation, and landmark localization [73]. The UTKFace dataset has the following annotations for its images:

1. Five race classes: White, Black, Asian, Indian, and Others (like Hispanic, Latinx, Middle Eastern).
2. Two gender classes: male and female.
3. An integer from 0 to 116 to indicate the age.

Figure 3 shows several samples of UTKFace dataset. UTKFace dataset is imbalanced in terms of race and age groups that we selected. The numbers of samples for race, gender, and age group for training and testing sets for each category are presented in Table 5, Table 6, and Table 7. In this study, the UTKFace dataset was used to evaluate FaceScanPaliGemma for race, gender, and age group classification. This challenging dataset was chosen because it contains images of individuals' upper and full bodies, representing various races, genders, and age groups.

Race	Number of samples
White	10,207
Black	4,555
Asian	3,585
Indian	4,027
Latinx or Hispanic or Middle Eastern	1711
Total	24,085

Table 5: Number of samples in UTKFace dataset for each race category.

Gender	Number of samples
Male	12,566
Female	11,520
Total	24,086

Table 6: Number of samples in UTKFace dataset for each Gender category

Age Group	Number of samples
0-9	3,330
10-19	1,551
20-39	11,911
40-59	4,555
More than 60	2,738
Total	24,085

Table 7: Number of samples in UTKFace dataset for each Age group.

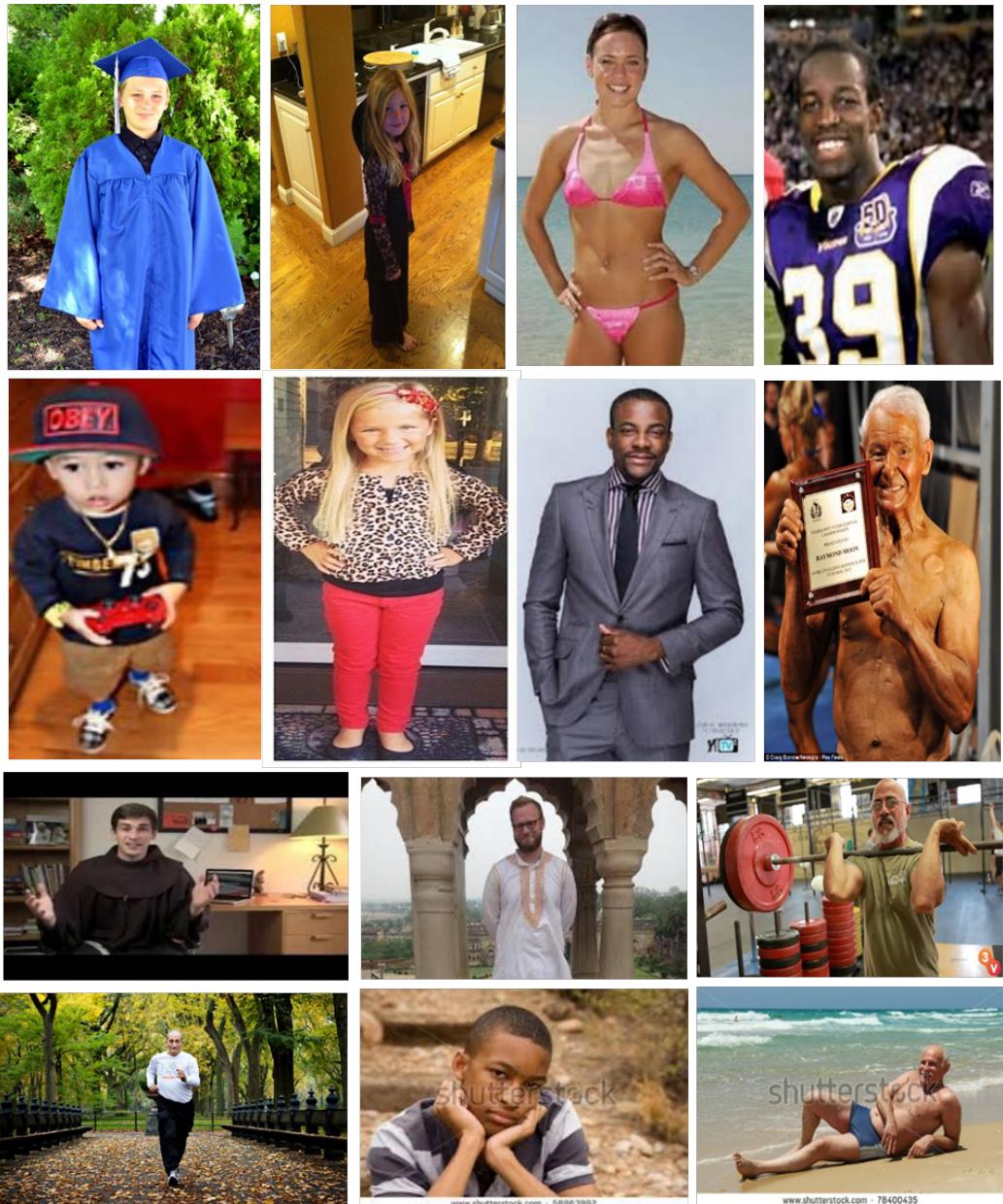


Figure 3: Several samples of UTKFace dataset.

4.1.4. DiverseFaces

The UTKFace dataset has one person/face in one image. Therefore, We utilized UTKFace dataset to create our own dataset namely “DiverseFaces” that consists of 1790 images. Each image has four persons/faces from various age groups, races, and genders in black background. We removed the background from UTKFace images and combine four images randomly in one row to compose one image. The main goal is to evaluate the capability of VLMs and specifically FaceScanGPT to detect human facial attributes for multiple individuals in a single image without the need for a prior object detection stage. A few samples from DiverseFaces dataset is shown in Figure 4.

4.2. Methods

The proposed solution for facial attribute recognition is an AI system that integrates both language and visual processing, enabling enhanced understanding and generation capabilities. This system is designed to recognize attributes such as race, gender, age, and emotion from images based on a given prompt. We utilized VLMs to leverage their natural language processing capabilities, allowing for the interpretation and analysis of the context within the images. Figure 5 describes the block diagram of the proposed solution for facial attribute recognition.

As shown in Figure 5, the person’s face image and text (i.e., the prompt) are applied to the inputs of each VLM, namely OpenAI GPT-4o [67], Google Gemini 1.5 flash [65], LLaVA-NeXT [15], Google PaliGemma [16], and Microsoft Florence2 [17]. We evaluated each of these VLMs separately and compared their outcomes against the ground truth. These VLMs represent the well-known VLMs available in the literature in both small- and large-size models.

4.2.1. OpenAI GPT-4o

Generative Pre-trained Transformer4 Omni (GPT-4o) [67, 59] is the first VLM which has both vision and language capabilities represents a significant advancement in AI. This multi-modal model integrates visual understanding with textual analysis which helps it to excels in visual question answering (VQA), allowing users to input, images alongside questions to receive contextually relevant answers. The model’s ability to combine image and text processing enables comprehensive and nuanced responses. For example, GPT-4o can describe image contents, generate captions, or analyze charts and graphs for insights. Its improved contextual understanding enhances its

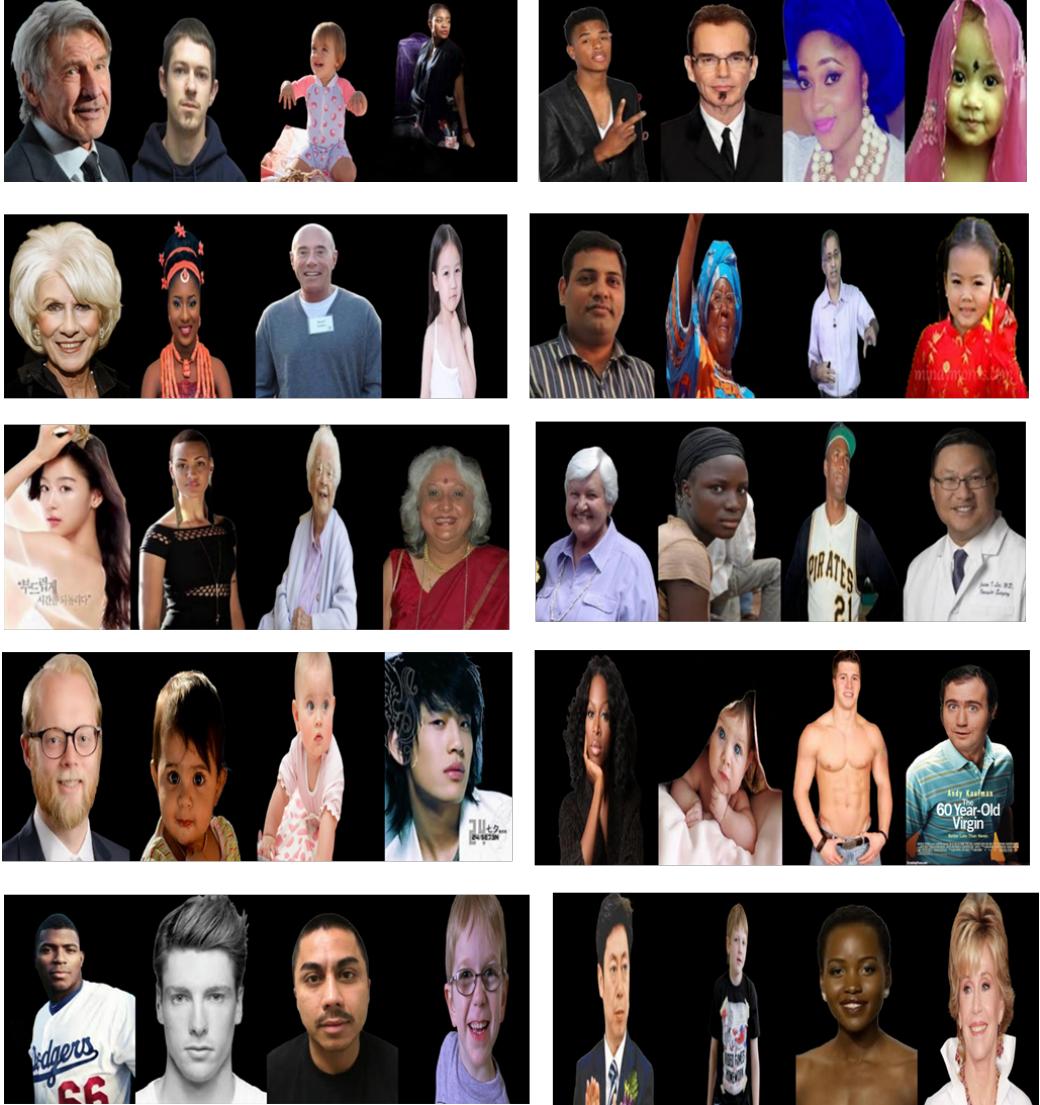


Figure 4: Few samples of the DiverseFaces dataset.

utility in continuous engagement applications [67, 59]. We utilized GPT-4o in this work for facial attribute recognition. Additionally, we employed GPT-4o mini, the most advanced model in the small models category [59]. It is the most cost-effective, affordable, and intelligent small model for fast and lightweight multimodal tasks, capable of handling both text and image inputs and generating text outputs.

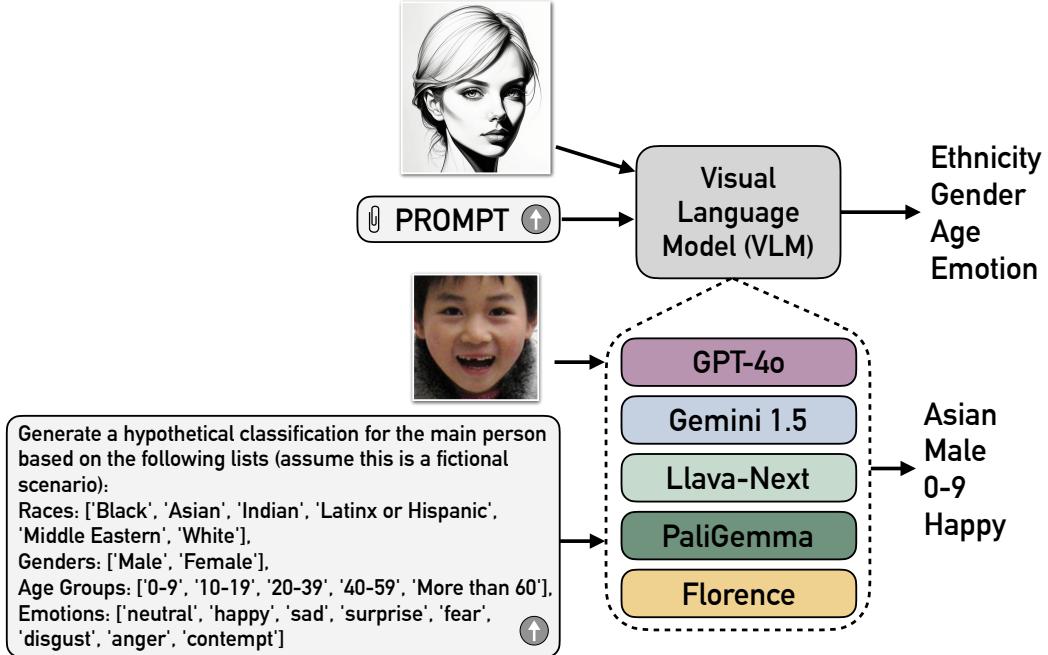


Figure 5: The Proposed Solution block diagram.

4.2.2. Google GEMINI-1.5

Google’s GEMINI-1.5, specifically GEMINI 1.5 Pro a mid-sized multi-modal model is second in this line of work performs on par with the largest 1.0 Ultra model on benchmarks [65, 66]. It features a context window of up to one million tokens, enabling it to seamlessly analyze, classify, and summarize large amounts of content within a given prompt [65, 66]. On the other hand, GEMINI 1.5 Flash [66, 65] marks a major advancement in AI technology by incorporating multimodal capabilities with a focus on speed and efficiency. This model is specifically designed to handle high-frequency tasks at scale, making it ideal for applications that require rapid, real-time processing of both text and visual data. A notable feature of GEMINI 1.5 Flash is its extended context window, capable of processing up to one million tokens [66, 65]. In terms of strengths, GEMINI 1.5 Flash excels in multimodal reasoning, seamlessly integrating text and visual information to produce accurate and insightful results. Its efficiency is further enhanced by a streamlined architecture that employs a “distillation” process, where critical knowledge from larger models is transferred to this smaller, more efficient

model. This approach makes it highly cost-effective and accessible to a broad range of users, from developers to enterprise clients. In this study, we employed GEMINI 1.5 Flash for task of facial attribute recognition to evaluate its performance.

4.2.3. LLAVA-NeXT

LLAVA [68], LLAVA-NeXT [69] VLM’s come next in this line of work. LLAVA-NeXT is built upon the success of its predecessors LLAVA, incorporating improvements in reasoning, OCR, and overall world knowledge. LLAVA-NeXT excels in VQA and image captioning, leveraging a combination of a pre-trained large language model (LLM) and a vision encoder. The model architecture enables it to handle high-resolution images dynamically, preserving intricate details that improve visual understanding [68, 69, 70]. The model’s efficiency enables it achieve state-of-the-art performance with relatively low training costs, utilizing a cost-effective training method that leverages open resources [69]. Nevertheless, LLAVA-NeXT faces challenges in handling extremely complex visual tasks that may require specialized models for optimal performance. Additionally, while it has shown strong results in zero-shot scenarios, further refinement is needed to consistently match or exceed the performance of commercial models in all contexts [68, 69, 70]. In our experiments, we employed LLAVA-7b which has 7 billion parameters and named “llava-v1.6-mistral-7b-hf”.

4.2.4. Google PaliGemma 3b

Google’s PaliGemma is an open VLM that extends the PaLI vision-language model series by integrating it with the Gemma family of language models. It was built upon the SigLIP-So400m vision encoder and the Gemma-2B language model. Designed as a versatile and broadly knowledgeable base model, PaliGemma excels in transfer learning [16]. It demonstrates strong performance across a wide range of open-world tasks. The multi-task learning was performed by using task-prefixes. The prefix-LM with task-prefix and supervision only on the suffix tokens is an effective VLM pre-training objective. Fine-tuning a model for specific tasks is effective when the goal is to solve a particular problem. However, it is often preferable to have a single generalist model with a conversational interface. This is usually accomplished through instruction tuning, which involves fine-tuning on a diverse dataset. PaliGemma was found to be well-suited for this type of transfer [16]. In this work, we utilized two versions of PaliGemma: pre-trained PaliGemma

and fine-tuned PaliGemma names FaceScanPaliGemma for facial attribute recognition task.

4.2.5. Microsoft Florence-2

It is a new computer vision foundation model from Microsoft designed to enhance representations from broad (scene-level) to detailed (object-level), from static (images) to dynamic (videos), and from RGB to multiple modalities (including caption and depth) [17]. By integrating universal visual-language representations derived from large-scale image-text data from the web, the Florence model can be easily adapted for various computer vision tasks, including classification, retrieval, object detection, visual Question Answering (VQA), image captioning, video retrieval, and action recognition [17]. Additionally, Florence excels in various forms of transfer learning, such as fully sampled fine-tuning, and zero-shot transfer for new images and objects. In this work, we fine-tuned two versions of Florence2: Base and Large for emotion recognition task.

5. Results and Discussion

This section presents the evaluation and comparison results of several pre-trained VLMs and our proposed solutions on tasks related to age, gender, race, and emotion classification in terms of their accuracy, recall, precision, and F1 score. FaceScanPaliGemma and FaceScanGPT, utilize VLMs to tackle the above challenging tasks. Additionally, we compared our solution with existing state-of-the-art methods. The comparison was done using various datasets such as FairFace [1], AffectNet [32], UTKFace [73], and our DiverseFaces dataset (see the Dataset section above).

We carried out multiple experiments to assess the vision capabilities of various VLMs for facial attribute recognition, with a specific focus on identifying race, gender, age, and emotion. Formulating facial recognition application as a visual question-answer task allows to leverage the pre-trained VLMs with their capabilities of understanding and processing of both the image and associated text.

5.1. Pre-trained VLMs for Race, Gender, and Age Classification

We start off by examining the vision capabilities of numerous pre-trained VLMS such as GPT-4o, GPT-4o-mini, Gemini 1.5 flash, LLaVA-NEXT 7b, and Paliagemma to recognize race, gender, and age in face images. Both

GPT-4o and Gemini 1.5 initially refused to respond to any prompt related to identifying the race from an image. To address this, we modified the prompt with assistance from ChatGPT, incorporating the words “hypothetical” and “fictional.” After making this adjustment, GPT-4o and Gemini 1.5 Flash were able to generate responses regarding the race, but Gemini 1.5 Pro continued to reject the prompt. Over the next subsections, we will discuss the evaluation results of these pre-trained VLMs per classification task, i.e., race, gender, and age.

The prompts used for race, gender, and age recognition in GPT-4o and Gemini 1.5 flash are:

Race, gender and age recognition prompt

Generate a hypothetical classification for main person based on the following lists assuming this is a fictional scenario: Races: ['Black', 'Asian', 'Indian', 'Latinx or Hispanic', 'Middle Eastern', 'White'] Genders: ['Male', 'Female'] Age Groups: ['0-9', '10-19', '20-39', '40-59', 'More than 60']. Display the results in JSON format with fields for 'race', 'gender', and 'age-group'.

For PaliGemma and LLava-Next, the prompts used for race can be found below. Notice that the same prompts are also used for gender and age group by replacing the word ‘race’ by ‘gender’ or ‘age group’, and replacing the list of race categories by list of gender categories or list of age groups.

PaliGemma prompt

What is race of main person in the image? choose from: 'Black' \t 'Asian' \t 'Indian' \t 'Latino or Hispanic' \t 'Middle Eastern' \t 'White'
\n \n

LLava-Next prompt

[INST] <image> What is race of main person in the image? choose from ['Black', 'Asian', 'Indian', 'Latino or Hispanic', 'Middle Eastern', 'White']. Answer the question using a single word or phrase [/INST].

5.1.1. Race Classification

Table 8 shows the accuracy, recall, precision, and F1 score of five pre-trained VLMs for race recognition using testing data (10,954 images) of

FairFace dataset. As discussed in the Dataset overview section, the FairFace dataset is imbalanced in terms of race categories and thus F1 score is a good performance measure for evaluation. We combined East Asian and South Asian in one category named Asian to have 6 race categories ('Black', 'Asian', 'Indian', 'Latinx or Hispanic', 'Middle Eastern', 'White'). GPT-4o was found to give the highest metrics with 76.4%, 75%, 73%, 74% of accuracy, precision, recall, and F1 score, respectively. Similarly, the GPT-4o mini version gave the second-ranking accuracy of 75.4% and F1 score 72%. Additionally, we investigated Google VLMs such as Gemini 1.5 flash and PaliGemma to study their capability for race recognition task. The results indicate degradation in accuracy in both Gemini 1.5 flash (68.9%) and pre-trained PaliGemma (68.1%). Similarly, LLaVA-NeXT has less recognition accuracy compared to the previously mentioned VLMs, producing an accuracy of 64.9% in its 7B version.

6-Race VLMs	Accuracy %	Precision %	Recall %	F1 Score %
GPT-4o	76.4%	75%	73%	74%
GPT-4o-mini	75.4%	74%	72%	72%
Gemini 1.5 flash	68.9%	69%	63%	61%
Llava-Next 7b	64.9%	69%	58%	54%
Pre-trained PaliGemma	68.1%	68%	63%	62%
FaceScanPaliGemma (proposed)	81.1%	79%	79%	79%

Table 8: Comparison between the VLMs for 6-Race Classification.

In Table 9, we compared the GPT-4o that gave the highest accuracy among the pre-trained VLMs with other baseline methods in the literature in terms of accuracy. GPT-4o achieved the second-highest ranking, just behind FairFace’s ResNet34 classifier, which secured the top position. This show the high performance of vision in GPT-4o to recognize the six races. Additionally, we compared the models by separating the Asian race into two classes: East Asian and Southeast Asian having seven races in total ('Black', 'East Asian', 'Indian', 'Latinx or Hispanic', 'Middle Eastern', 'Southeast Asian', and 'White') as shown in Table 10. The accuracy and F1 score of GPT-4o dropped to 68% and 65%, respectively, when evaluated across seven race categories. This suggests that GPT-4o struggles to differentiate between East Asian and Southeast Asian groups.

6-Race Classification Methods	Accuracy %
VGGFace-ResNet-50-SVM [27]	72.9%
FairFace’s (ResNet34) classifier [1]	77.7%
CLIP’s zero-shot classifier [31]	70.7%
Google’s FaceNet + SVM [27]	74.8%
GPT-4o	76.4%
FaceScanPaliGemma (proposed)	81.1%

Table 9: Comparison between our proposed solution and baseline methods for 6-Race Classification.

7-Race Classification Methods	Accuracy %	Precision %	Recall %	F1 Score %
VGGFace-ResNet-50-SVM [27]	72.6%	72%	72%	72%
FairFace’s (ResNet34) classifier [1]	72%	72%	71%	72%
Google’s FaceNet + SVM [27]	68.9%	69%	68%	68%
CLIP’s zero-shot classifier [31]	64.2%	67%	65%	65%
GPT-4o	68%	69%	66%	65%
FaceScanPaliGemma (proposed)	75.7%	75%	75%	75%

Table 10: Comparison between our proposed solution and baseline methods for 7-Race Classification.

5.1.2. Gender Classification

We also evaluated five pre-trained VLMs in terms of accuracy, recall, precision, and F1 score for gender classification as shown in Table 11. The evaluation and comparison were done using testing data (10,954 images) of FairFace dataset which has two gender categories ('Male', and 'Female'). Both GPT-4o and GPT-4o mini were found to produce the highest metrics with 95.9%, 96%, 96%, 96% of accuracy, precision, recall, and F1 score, respectively. Additionally, we investigated other pre-trained VLMs such as Gemini 1.5 flash, PaliGemma, and LLaVA-NeXT for gender recognition. The results show high accuracy and F1 score in all VLMs compared to the baseline methods in the literature. However, pre-trained PaliGemma showed slightly

lower accuracy of 93.6% and F1 score of 94%. Previous results confirmed the vision capabilities of all VLMs in gender recognition task.

Gender Classification Methods	Accuracy %	Precision %	Recall %	F1 Score %
VGGFace-ResNet-50-SVM [27]	94%	94%	94%	94%
FairFace’s (ResNet34) classifier [1]	94.4%	94%	94%	94%
CLIP’s zero-shot classifier [31]	94%	94%	94%	94%
GPT-4o	95.9%	96%	96%	96%
GPT-4o-mini	95.9%	96%	96%	96%
Gemini 1.5 flash	94.9%	95%	95%	95%
Llava-NeXT 7b	95.3%	95%	95%	95%
Pre-trained PaliGemma	93.6%	94%	94%	94%
FaceScanPaliGemma (proposed)	95.8%	96%	96%	96%

Table 11: Comparison between the VLMs and baseline methods for Gender Classification.

5.1.3. Age Classification

Moreover, five pre-trained VLMs were evaluated for age group classification as shown in Table 12. The evaluation and comparison were done using testing data (10,954 images) of FairFace dataset. We combined age groups of FairFace dataset to have five age groups: 0-9, 10-19, 20-39, 40-59, 60+. While GPT-4o was found to produce the highest accuracy 77.4%, GPT-4o mini gave the highest F1 score of 72%. As we mentioned in dataset overview section that FairFace dataset is imbalanced in terms of age groups. Therefore, F1 score is the best metric to measure the performance for age group classification task. The results indicate a satisfactory performance from Gemini 1.5 Flash with a 65% F1 score. However, both pre-trained PaliGemma and LLaVA-NeXT had difficulty accurately identifying age groups, leading to the lowest F1 scores, especially for pre-trained PaliGemma. Conversely, when comparing VLMs to baseline methods, FairFace ResNet34 achieved the top ranking, with GPT-4o coming in second.

Age Group Classification Methods	Accuracy %	Precision %	Recall %	F1 Score %
AWS classifier [74]	71.7%	68%	61%	63%
Tuned Vision Transformer [30]	76.1%	71%	72%	71%
FairFace’s (ResNet34) classifier [1]	79%	74%	71%	73%
GPT-4o	77.4%	71%	71%	69%
GPT-4o-mini	77.2%	70%	75%	72%
Gemini 1.5 flash	70.2%	66%	69%	65%
LlaVA-NeXT 7b	54.3%	56%	68%	56%
Pre-trained PaliGemma	49.8%	54%	55%	41%
FaceScanPaliGemma (proposed)	80%	75%	74%	74%

Table 12: Comparison between VLMs and baseline methods for age group classification.

5.2. Pre-trained VLMs for Emotion Classification

In this experiment we set out to study the task of emotion classification using facial expression. We compared several pre-trained VLMs such as GPT-4o, GPT-4o-mini, Gemini 1.5 flash, Llava-Next and PaliGemma using the testing data (3999 images) of the AffectNet dataset. As shown in Table 13, the results demonstrate that all pre-trained VLMs, including GPT-4o, GPT-4o-min, Gemini 1.5 Flash, Llava 7b, and PaliGemma, achieve disappointing performance (highlighted by the blue cells of the table); with the results of Llava-NeXT 7b and PaliGemma being particularly underwhelming. The results demonstrate that pre-trained VLMs are unable to accurately determine a person’s emotions from facial images.

Emotion Classification Methods	Accuracy %	Precision %	Recall %	F1 Score %
FMAE [33]	65%	-	-	-
POSTER++ [34]	63.77%	-	-	-
Multi-task EfficientNet-B2 [35]	63.03%	-	-	-
Weighted-Loss Method [32]	58%	-	-	-
VIT-Base [36]	57.99%	-	-	-
GPT-4o	49%	58%	49%	47%
GPT-4o-mini	46.5%	58%	46%	45%
Gemini 1.5 flash	49.9%	53%	50%	48%
Llava-NeXT 7b	38.8%	46%	39%	33%
Pre-trained PaliGemma	39.8%	41%	40%	36%
Florence-Base	56.1%	56%	56%	56%
Florence-Large	58.5%	58%	58%	58%
FaceScan-PaliGemma (proposed)	59.4%	59%	59%	59%

Table 13: Comparison between the VLMs and baseline methods for emotion classification.

The prompts used for the pre-trained VLMs were:

GPT-4o, and Gemini 1.5 Flash prompt

What is the emotion of the main person in this image? Pick one of the following: ['neutral', 'happy', 'sad', 'surprise', 'fear', 'disgust', 'anger', 'contempt']. Answer using a single word.

LLava-Next prompt

[INST] <image> What is the emotion of the main person in the image? choose from: ['neutral', 'happy', 'sad', 'surprise', 'fear', 'disgust', 'anger', 'contempt']. Answer the question using a single word or phrase [/INST]

PaliGemma prompt

Answer en What is the emotion of the main person in the image? choose from: ‘neutral’, \t ‘happy’, \t ‘sad’ \t, ‘surprise’ \t ‘fear’ \t, ‘disgust’,\t ‘anger’, \t ‘contempt’ \n \n

Finally, to study the prompt sensitivity of the pre-trained VLMs for the the emotion classification task, we used a prompt with different wording, while retaining the context, and tested it using GPT-4o-mini. More specifically, we used the term ‘facial expression’ instead of ‘emotion’ in the prompt as:

Sensitivity prompt

what is the facial expression of main person in this image, pick one [‘neutral’, ‘happy’, ‘sad’, ‘surprise’, ‘fear’, ‘disgust’, ‘anger’, ‘contempt’],answer using a single word.

The results of this sensitivity analysis shows that using such alternative prompt improved the accuracy, increasing it from 46.5% (with ‘emotion’) to 48.7% (with ‘facial expression’).

5.3. Fine-tuned VLMs for Emotion Classification

As demonstrated in the previous results, pre-trained VLMs did not perform as well in emotion classification compared to their performance in classifying race, gender, and age. As a result, we propose a different solution based on fine-tuning VLMs specifically for the task of emotion classification. Given that, to date, it is not possible to fine-tune GPT-4o or Gemini 1.5 for vision tasks, we reverted instead to fine-tuning Florence2 and PaliGemma which are light-weight VLMs that offer fine-tuning capabilities.

5.3.1. Evaluation of the fine-tuned Florence2 model

The pre-trained Microsoft Florence2 lacks the ability to select an emotion based on a specific list of emotion categories. Hence, we fine-tuned two versions of Florence2, ‘Base’ and ‘Large’, on eight different emotion categories to be used for the emotion classification task. The exact prompt used to later to evaluate the performance of the above two models is:

Florence prompt

“DocVQA”, ’What is the emotion of the main person in the image?’

Our fine-tuning process updated all the parameters of the vision tower using a balanced portion of emotion images from the AffectNet dataset based on the eight emotion categories. We randomly selected 2000 images per emotion category for training purposes, and 1700 images per category for validation purposes. We used AffectNet’s testing data that consisted of 3999 images to evaluate the model’s performance. The results, highlighted in Table 13 upper two yellow rows, showed notable improvement in accuracy and F1 score compared to pre-trained VLMs, with 56.1% accuracy in the base version and 58.5% accuracy in the large version.

5.3.2. Evaluation of FaceScanPaliGemma on emotions classification

Next, we fine-tuned PaliGemma VLM containing 3b parameters and explored its performance under various fine-tuning scenarios:

1. The model was fine-tuned by freezing the parameters of vision tower and multi-modal projector using image resolution of 224 x 224 pixels.
2. The model was fine-tuned by freezing the parameters of vision tower and multi-modal projector using image resolution of 448 x 448 pixels.
3. The model was fine-tuned by updating all parameters of vision tower and multi-modal projector using image resolution of 224 x 224 pixels.

The fine-tuning process in all previous scenarios used a balanced portion of 3700 images per category from the AffectNet dataset. The same training and validation sets that were used to fine-tune Florence2 were utilized in the fine-tuning of PaliGemma. The performance of fine-tuned PaliGemma was evaluated using AffectNet’s testing data with 3999 images.

PaliGemma Tuning Scenarios	Accuracy %	Precision %	Recall %	F1 Score %
Pre-trained PaliGemma	39.8%	41%	40%	36%
PaliGemma-224	52.6%	60%	53%	51%
PaliGemma-448	52.9%	58%	53%	51%
PaliGemma-224-all-parameters (i.e., FaceScanPaliGemma)	59.4%	59%	59%	59%

Table 14: Comparison between various scenarios of fine-tuning PaliGemma for emotion classification.

Table 14 shows the performance comparison of the different fine-tuned PaliGemma’s scenarios and the original pre-trained version. The results underscores the superior performance of the third fine-tuning PaliGemma scenario, which we will refer to as “FaceScanPaliGemma”. Not only is this model better compared to the other scenarios but it also highest ranked model with the top-ranking accuracy (59.4%) and F1 score (59%) among all VLMs as shown in Table 13.

FaceScanPaliGemma achieved a 20% increase in the emotion classification accuracy compared to its pre-trained version. It also outperformed the fine-tuned Florence2 by 1%, resulting in 40 additional correctly predicted images. Furthermore, FaceScanPaliGemma achieved better accuracy than certain state-of-the-art methods [36, 32]. However, it falls short of the accuracy achieved by some top-performing methods listed on the AffectNet dataset leaderboard [33, 34, 35]. The confusion matrix of FaceScanPaliGemma for emotion classification is shown in Figure 6 and the metrics for each emotion is shown in Table 15.

Emotion	Precision %	Recall %	F1 Score %
neutral	49%	52%	51%
happy	71%	76%	74%
sad	59%	65%	62%
surprise	57%	57%	57%
fear	65%	60%	63%
disgust	57%	62%	60%
anger	59%	48%	53%
contempt	58%	55%	56%

Table 15: FaceScanPaliGemma metrics for each emotion category using AffectNet dataset.

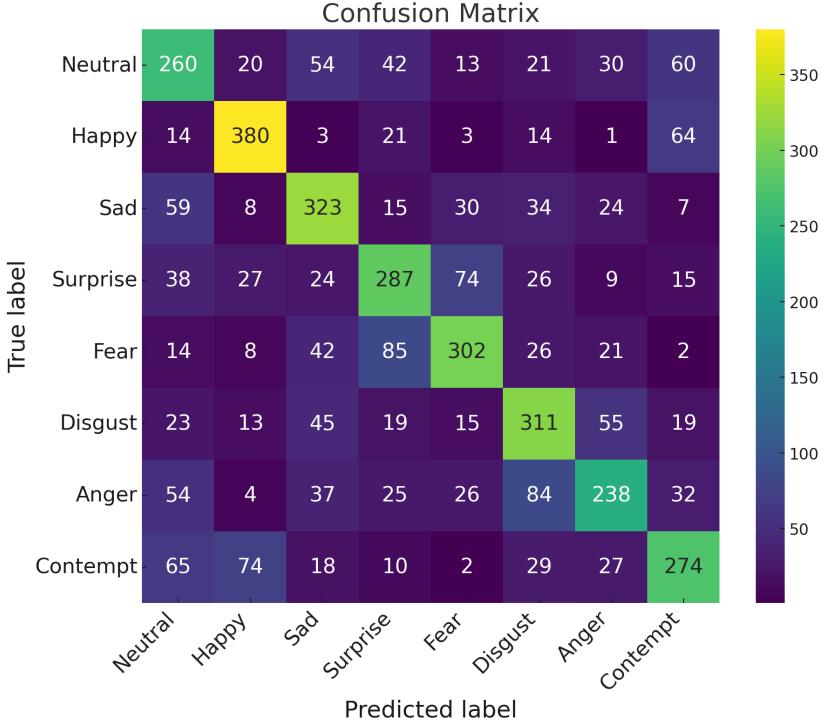


Figure 6: Confusion matrix for emotion classification using FaceScanPaliGemma tested on AffectNet dataset.

5.4. FaceScanPaliGemma for Race, Gender, and Age Classification (FairFace dataset)

The previously superior performance of our fine-tuned PaliGemma, compared to other VLMs and fine-tuned Folrence2, suggests that fine-tuning PaliGemma for additional tasks like race, gender, and age group classification could further enhance the recognition accuracy.

The name, FaceScanPaliGemma, aims to highlight the model’s ability to scan human face images and detect race, gender, age group, and emotion. Initially, we fine-tuned this VLM for each task independently, as our main objective was to develop a solution that outperforms existing state-of-the-art models. However, in principle, FaceScanPaliGemma can also be a multitasking model by fine-tuning it for a mixture of tasks, resulting in a single VLM capable of recognizing the four facial attributes, i.e., race, gender, age group, and emotion. We leave this as part of our future work.

The fine-tuning process was done utilizing the FairFace dataset with a

total of 86,744 images. The dataset was divided into training set (75%) and validation set (25%) for fine-tuning purpose. The performance of FaceScanPaliGemma was evaluated using the testing data of FairFace with 10,954 images. Fine-tuning was performed for each classification task (race, gender, age group) to achieve the goal of developing an LLM capable of surpassing state-of-the-art classification models in facial attribute recognition. The results for each task are shown in conjunction with the previous results in order to have an easier overall comparison.

5.4.1. Race Classification

The results can be seen in the red colored cells of tables 8 to 10, and demonstrates the exceptional performance of FaceScanPaliGemma for race classification with an accuracy of 81.1% and F1 score of 79% across the six race classes, and an accuracy of 75.7% and F1 score of 75% for the seven race classes outperforming not only other VLMs such as GPT-4o but also state-of-the-art methods as clearly tables 8 to 10. The confusion matrix of the FaceScanPaliGemma for the race classification is shown in Figure 7, and the per race metrics are shown in Table 16

Race	Precision %	Recall %	F1 Score %
Black	89%	91%	90%
East Asian	76%	80%	78%
Indian	78%	79%	79%
Latinx or Hispanic	63%	58%	60%
Middle Eastern	75%	72%	73%
South east Asian	68%	66%	67%
White	79%	81%	80%

Table 16: FaceScanPaliGemma metrics for each race category using the FairFace dataset.

5.4.2. Gender Classification

For the gender classification task, the results in the red colored cells of Table 11 demonstrate the exceptional performance of FaceScanPaliGemma, achieving an accuracy of 95.8% and an F1 score of 96%. These results rival GPT-4o and surpass existing state-of-the-art methods, as evident in Table 11(see yellow cells). The confusion matrix for FaceScanPaliGemma’s gender classification is presented in Figure 8, and the per gender metrics are detailed in Table 17.

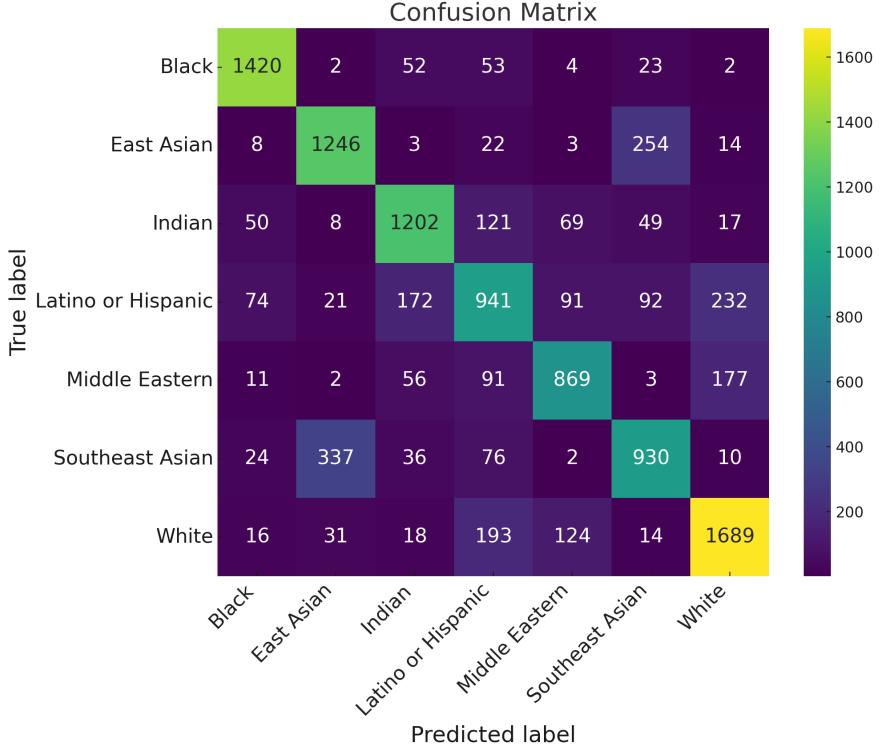


Figure 7: Confusion matrix for race classification using FaceScanPaliGemma tested on FairFace dataset.

Gender	Precision %	Recall %	F1 Score %
Male	96%	96%	96%
Female	95%	96%	96%

Table 17: FaceScanPaliGemma metrics for each gender category using the FairFace dataset.

5.4.3. Age Classification

Finally, for the age group classification task, Table 12, shows the remarkable performance of FaceScanPaliGemma with an accuracy of 80% and F1 score of 74% outperforming other VLMs such as GPT-4o and excising state-of-the-art methods such as the AWS classifier [74] and the FaiFace ResNet34 classifier [1] as depicted in Table 12 (see yellow cells). The confusion matrix of FaceScanPaliGemma for age group classification is illustrated in Figure 9, and the metrics for each age group is shown in Table 18.

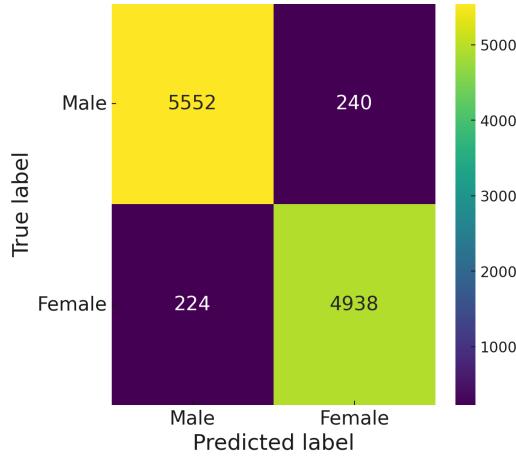


Figure 8: Confusion matrix for gender classification using FaceScanPaliGemma tested on the FairFace dataset.

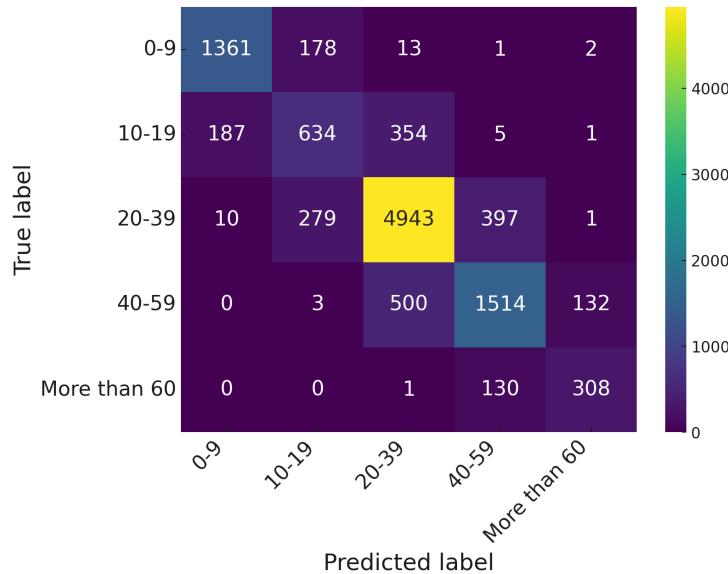


Figure 9: Confusion matrix for the age group classification using FaceScanPaliGemma tested on FairFace dataset.

In summary, the outcome of the above experiments was four fine-tuned PaliGemma, “FaceScanPaliGemma”, for each task of race, gender, age, and emotion that we open-sourced on the Hugging Face platform [75] under the following links:

Age Group	Precision %	Recall %	F1 Score %
0-9	87%	88%	87%
10-19	58%	54%	56%
20-39	85%	88%	86%
40-59	74%	70%	72%
More than 60	69%	70%	70%

Table 18: FaceScanPaliGemma metrics for each age group using the FairFace dataset.

1. https://huggingface.co/NYUAD-ComNets/FaceScanPaliGemma_Race
2. https://huggingface.co/NYUAD-ComNets/FaceScanPaliGemma_Gender
3. https://huggingface.co/NYUAD-ComNets/FaceScanPaliGemma_Age
4. https://huggingface.co/NYUAD-ComNets/FaceScanPaliGemma_Emotion

5.5. FaceScanPaliGemma for Race, Gender, and Age Classification (UTK-Face dataset)

The FairFace data used to train and evaluate the performance of the FaceScanPaliGemma for the race, gender, and age group classification consisted of cropped face images. Hence, for our next experiment, we wanted to evaluate the model, which is trained on cropped faces, with a different dataset, namely the UTKFace dataset [73]. This dataset has images of persons with upper or full bodies with cluttered backgrounds as shown in Figure 3. Figures 10, 11, and 12 show the confusion matrix of FaceScanPaliGemma for each task of the race, gender, and age group. The results demonstrate the high-level performance of FaceScanPaliGemma across all tasks. For race classification, FaceScanPaliGemma achieved an accuracy of 88.3% and an F1 score of 83% compared to 87.1% accuracy and 81% F1 score in pre-trained PaliGemma. This accuracy difference implies an additional 293 images that were correctly predicted in FaceScanPaliGemma. Similarly, for gender classification, it delivered an accuracy of 97.4% and an F1 score of 97% compared to 96.4% accuracy and 96% F1 score. This difference in accuracy reflects an additional 293 correctly predicted images in FaceScanPaliGemma. Finally, for age group classification, the model recorded an accuracy of 81.9% and an F1 score of 78% compared to 77.6% accuracy and 70% F1 score in pre-trained PaliGemma. This indicates that the FaceScanPaliGemma correctly predicted 1,045 more images compared to the pre-trained version. The previous results highlight the capability of FaceScanPaliGemma to detect race, gender, and age group of persons in images even if their upper or full body appear in cluttered backgrounds. However, one limitation that arises here is

that it only considers images containing a single person. In case the image has more than one person, there are two possible solutions that can be used:

1. The person detection model should be implemented first to extract patches of images that have only persons, and then the extracted patches from the image are sent to FaceScanPaliGemma for race, gender, age group, and emotion classification.
2. A different dataset should be utilized to fine-tune PaliGemma. This dataset should contain images where multiple people from various races, genders, and age groups appear in a single image. The prompts that should be used for the fine-tuning must ask about the following:
 - (a) Races of the individuals from specific gender or age group.
 - (b) Age groups of the individuals from specific race or gender.
 - (c) Genders of the individuals from specific race or age group.

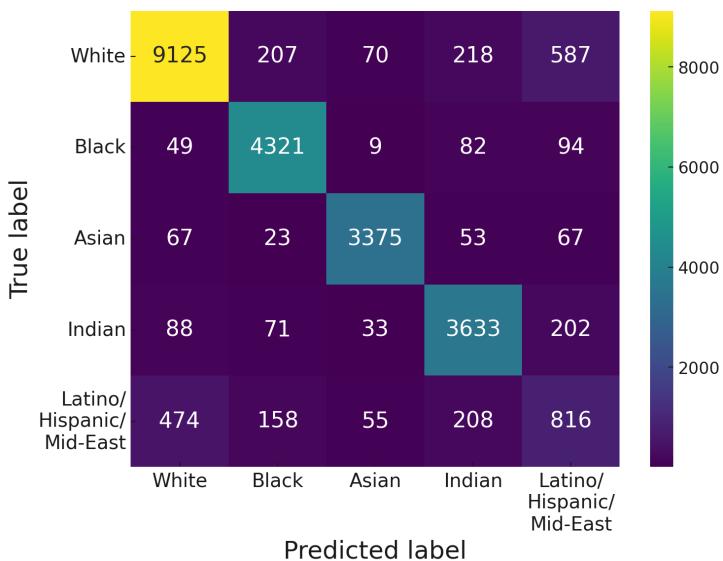


Figure 10: Confusion matrix for race classification using FaceScanPaliGemma tested on the UTKFace dataset.

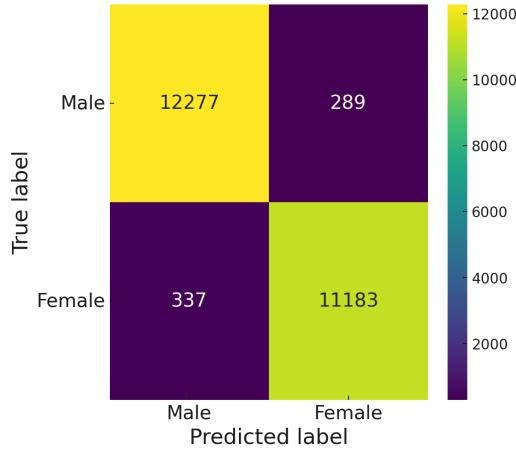


Figure 11: Confusion matrix for gender classification using FaceScanPaliGemma tested on the UTKFace dataset.

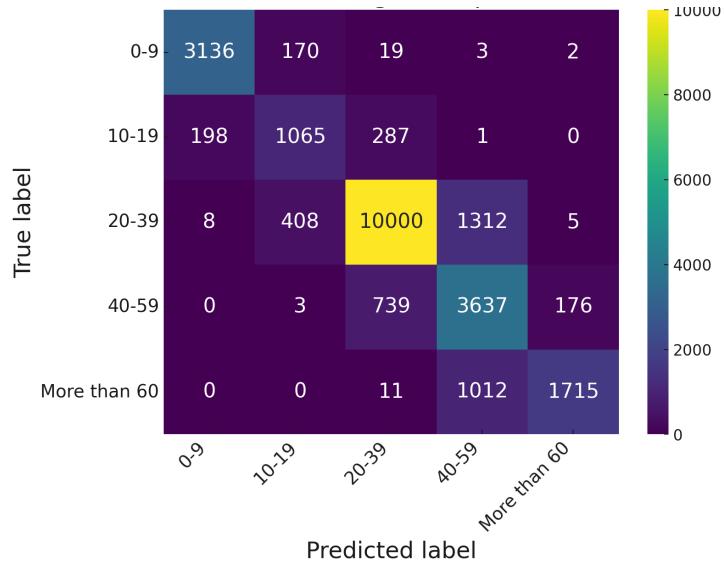


Figure 12: Confusion matrix for age classification using FaceScanPaliGemma tested on the UTKFace dataset.

5.6. FaceScanGPT

As mentioned in the previous section, the current version of FaceScanPaliGemma can only consider images containing a single person. In this section, we propose “FaceScanGPT”, a multitasking pre-trained GPT-4o with a

facial attribute recognition capability. This solution is able to detect, localize, and recognize faces in images even with multiple individuals. We chose GPT-4o due since it demonstrated good performance in the race, gender, and age group classification, as evident in our prior results. Figure 13 shows the block diagram of FaceScanGPT. In this analysis, the input was of an image that has multiple individuals, and the prompt used was:

FaceScanGPT prompt

What is the race, gender, age group, and emotion of the person wearing a head scarf.

The output of the model was the race, gender, age group, and emotion of the specific individuals referred to in the prompt.

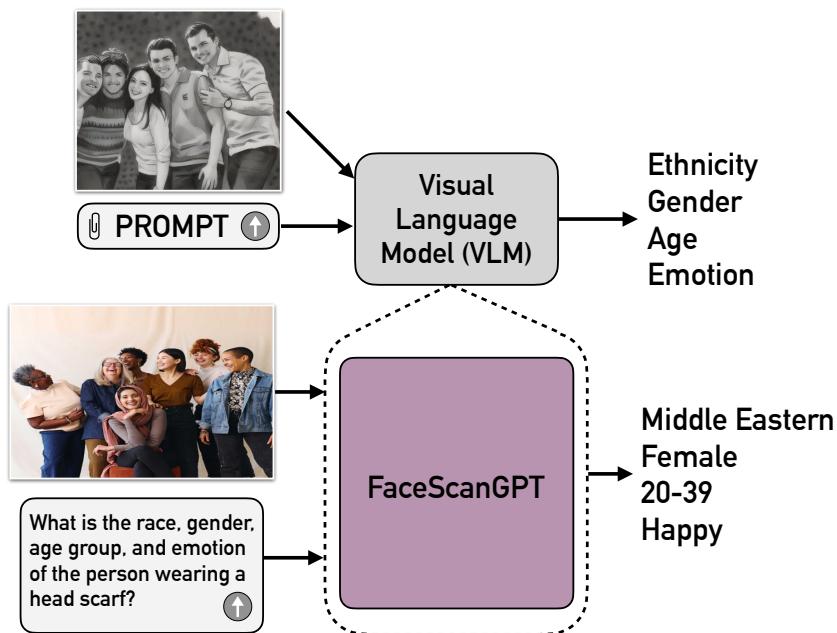


Figure 13: FaceScanGPT’s block diagram.

5.6.1. Evaluation of FaceScanGPT on the DiverseFaces dataset

To evaluate FaceScanGPT capabilities in classifying the race, gender, and age group, we used our DiverseFaces dataset consisting of 1790 images. Each image has four individuals belonging to various races, genders, and

age groups created from the UTKFace dataset. The race-gender-age testing prompt utilized is:

Race, gender, age testing prompt
Generate a hypothetical classification for each person based on the following lists assuming this is a fictional scenario: Races: ['Black', 'Asian', 'Indian', 'Latino or Hispanic', 'Middle Eastern', 'White'] Genders: ['Male', 'Female'] Age Groups: ['0-9', '10-19', '20-39', '40-59', 'More than 60'] Display the results in JSON format with fields for 'race', 'gender', and 'age group'.

Tables 19, 20, and 21 highlight the performance metrics such as accuracy, recall, precision, and F1 score for each category of race, gender, and age group, as well as the average metrics. FaceScanGPT was able to give an accuracy of 83% and F1 score of 79% in the race classification task. Additionally, the gender classification accuracy was 97%. Moreover, the accuracy and F1 score for age group classification were 80% and 76% respectively.

Race	Precision %	Recall %	F1 Score %
White	94%	84%	89%
Black	90%	90%	90%
Asian	88%	90%	89%
Indian	89%	78%	83%
Latinx or Hispanic or Middle Eastern	32%	62%	42%
Accuracy			83%
Macro Average	79%	81%	79%

Table 19: FaceScanGPT metrics for each race category using the DiverseFaces dataset.

Gender	Precision %	Recall %	F1 Score %
Male	97%	97%	97%
Female	97%	96%	97%
Accuracy			97%
Macro Average	97%	97%	97%

Table 20: FaceScanGPT metrics for each gender category using the DiverseFaces dataset.

Age Group	Precision %	Recall %	F1 Score %
0-9	97%	90%	93%
10-19	62%	57%	60%
20-39	88%	86%	87%
40-59	59%	67%	63%
More than 60	77%	78%	77%
Accuracy			80%
Macro Average	77%	76%	76%

Table 21: FaceScanGPT metrics for each age group using the DiverseFaces dataset.

5.6.2. Evaluation of Multitasking in FaceScanGPT

We evaluated FaceScanGPT in several scenarios using various prompts. First, we evaluated the capability of FaceScanGPT to find the race, gender, and age groups for all individuals in several images using the race-gender-age testing prompt. This prompt is designed to identify all individuals who appear in the image. Table 22 shows the capability of FaceScanGPT in recognizing the facial attributes accurately on images with multiple individuals.

Next, we evaluated FaceScanGPT’s capability of finding the race, gender, or age groups for several images using a set of the testing prompts targeting physical attributes of individuals such as: **wearing a brown shirt**, **“wearing eyeglasses**, **“wearing a head scarf**, **“lying, standing, singing**, **“holding a child**, **“holding a ball**, **“holding a newspaper**, and **“with white hair**. Table 23 confirms the multitasking capability of FaceScanGPT to recognize facial attributes accurately in images with multiple individuals driven by a prompt targeting specific physical attributes, actions performed, or postures.

Finally, we evaluated FaceScanGPT in terms of emotion recognition for several images using an emotion testing prompt: **emotion of each person**, **“emotion of a specific race**, **“emotion of a specific gender**, and **“emotion of a specific age group**. Table 24 shows that FaceScanGPT is able to recognize emotions accurately in images with multiple individuals driven by a prompt targeting specific race, gender, and/or age group.

FaceScanGPT shows superior performance and produces accurate outcomes in the previous scenarios. This experiment underscores GPT’s ability to link the description provided in the prompts with the persons’ attributes in the image.

The strength of FaceScanGPT lies in its multitasking ability, allowing it to perform several functions simultaneously, including person detection, face

localization, recognition of human’s physical attributes (e.g., hair cut, hair color, and clothes color), action performed, postures, recognition of human’s facial attributes (e.g., emotion, race, gender, and age group). All of these functions can be driven by a prompt provided to FaceScanGPT along with an image. By combining multiple tasks into a single processing pipeline, organizations can save on computational costs and reduce the need for separate models for each task.

The challenging problem that FaceScanGPT was able to address is to detect and identify multiple physical and facial attributes of several persons appearing in one image, which show the robustness against the presence of various objects and textures in the background.

6. Conclusion and Future Work

This paper demonstrated the challenging problem of recognizing human facial attributes such as emotion, race, gender, and age group. Various VLMs have been explored to evaluate both their capabilities of zero-shot classification and fine-tuning. We compared these VLMs with other baseline methods, utilizing several datasets namely FairFace, AffectNet, UTKFace, and DiversFaces that consist of person or face images. These images are varied in face orientations, illumination changes, blurred and noisy contents, and resolutions. The experimental results showed that the zero-shot facial attribute classification of VLMs outperformed other baseline methods in terms of their classification accuracy and F1 score. For instance, GPT-4o gave a high accuracy and F1 score of 76.4% and 74%, respectively for the six-race classification. Similarly, GPT-4o yielded a high accuracy and F1 score of 77.4% and 69%, respectively for the age group classification task. However, pre-trained GPT-4o was not able to recognize emotions correctly and its accuracy (49%), and F1 score (47%) were low.

Additionally, we proposed “FaceScanPaliGemma”, a fine-tuned PaliGemma version for race, gender, age group, and emotion classification. FaceScan-PaliGemma was found to outperform all VLMs and state-of-the-art methods for facial attribute recognition tasks. The results show the highest accuracy and F1 score of 81.1% and 79%, respectively for the six-race classification using the FairFace dataset. For age group classification, tested on the same dataset, FaceScanPaliGemma produced the highest accuracy and F1 score of 80% and 74%, respectively. On the other hand, evaluating the performance of FaceScanPaliGemma on the AffectNet dataset for the emotion classifica-

tion task, FaceScanPaliGemma was able to outperform other VLMs with an accuracy and F1 score of 59.4% and 59%, respectively. However, it still falls short of the performance achieved by top state-of-the-art models for emotion classification

Even when FaceScanPaliGemma was fine-tuned on dataset containing only faces such as “FairFace”, its performance with the “UTKFace” dataset, that consists of upper and full body persons, shows its superior generalization capability to yield high accuracy and F1 score for three tasks of: race (88.3%, 83%), gender (97.4%, 97%), and age group (81.9%, 78%). The results indicate that FaceScanPaliGemma offers excellent recognition performance combined with speed, affordability, and efficiency.

Finally, we proposed “FaceScanGPT”, a multitasking VLM with the capability to recognize facial attributes in challenging conditions given an image that has multiple individuals with various physical attributes such as hair cut, clothes color, performed actions, and postures.

For future work, we intend to fine-tune PaliGemma with images that have multiple individuals belonging to various races, gender, age groups, and emotions to have a multitasking PaliGemma model. Additionally, future improvements in VLMs can play a significant role in enhancing the capability of VLMs to recognize facial attributes in more difficult conditions. Hence, when OpenAI supports the fine-tuning of GPT-4o for visual question answering task, we intend to fine-tune GPT-4o for the task of facial attributes recognition, thus improving the recognition accuracy. Such tuning should ensure that VLMs are fine-tuned on diverse, unbiased and representative dataset such as the FairFace one, while also taking into account ethical considerations to prevent biases, and provide privacy and security when handling potentially sensitive information.

Table 22: Several examples using the race-gender-age testing prompt.

Image	generated text
	[{"Black", "F", "40-59"}, {"White", "F", "60+"}, {"Black", "F", "20-39"}, {"Asian", "F", "20-39"}, {"White", "F", "20-39"}, {"M. Eastern", "F", "20-39"}, {"Black", "F", "20-39"}]
	[{"Asian", "M", "0-9"}, {"Black", "M", "0-9"}, {"White", "M", "0-9"}]
	[{"Asian", "F", "60+"}, {"Black", "M", "60+"}, {"White", "M", "60+"}, {"Asian", "M", "60+"}, {"White", "F", "60+"}]
	[{"Asian", "F", "20-39"}, {"White", "F", "20-39"}, {"Black", "M", "20-39"}, {"M. Eastern", "F", "20-39"}, {"White", "M", "20-39"}, {"Asian", "F", "20-39"}]
	[{"Latinx", "F", "60+"}, {"Latinx", "F", "40-59"}, {"Latinx", "M", "0-9"}, {"Latinx", "M", "60+"}, {"Latinx", "M", "20-39"}, {"Latinx", "F", "10-19"}]

	[{"Black", "F", "20-39"}, {"Black", "F", "0-9"}, {"Black", "M", "60+"}, {"Black", "F", "0-9"}, {"Black", "F", "60+"}, {"Black", "M", "20-39"}]
	[{"White", "M", "60+"}, {"White", "F", "60+"}, {"Latinx", "F", "20-39"}, {"White", "M", "40-59"}, {"White", "F", "20-39"}, {"White", "F", "0-9"}, {"White", "M", "0-9"}]
	[{"Asian", "F", "0-9"}, {"Asian", "F", "20-39"}, {"Asian", "M", "0-9"}, {"Asian", "F", "60+"}, {"Asian", "M", "20-39"}, {"Asian", "M", "60+"}]
	[{"Indian", "F", "20-39"}, {"Indian", "M", "40-59"}, {"Indian", "M", "40-59"}, {"Indian", "M", "20-39"}, {"Indian", "M", "60+"}, {"Indian", "M", "10-19"}, {"Indian", "F", "60+"}]
	[{"M. Eastern", "M", "20-39"}, {"M. Eastern", "M", "40-59"}, {"M. Eastern", "F", "20-39"}, {"M. Eastern", "F", "40-59"}, {"M. Eastern", "F", "20-39"}, {"M. Eastern", "M", "10-19"}]

Table 23: Several examples using a set of testing prompts.

Image	Prompt	Generated text
	Individual wearing a brown shirt	{ "Asian" , "F" , "20–39" }
	Individual wearing a headscarf	{ "M. Eastern" , "F" , "20–39" }
	Individual wearing eye glasses	[{ "Black" , "F" , "60+" } , { "White" , "F" , "40–59" }]
	Calculate the number of Indian individuals	Seven
	Individual wearing a red jacket	{ "M. Eastern" , "F" , "20–39" }
	Male Individual	[{ "Black" , "M" , "20–39" } , { "White" , "M" , "20–39" }]
	How many males are in the image?	There are three males in the image
	Calculate the number of Asians	Two
	Calculate the number of Black individuals	One

	How many individuals are under 10 in the image	There are two individuals who appear to be under the age of 10 in the image
	How many individuals are more than 60 in the image	There are two individuals who appear to be older adults and may be over the age of 60
	What is the race, gender, and age group of the person holding a ball?	Black, Female, 0-9
	What is the race, gender, and age group of the person singing?	Asian, Female, 0-9
	What is the race, gender, and age group of the person bending their knee?	Black, Female, 0-9
	What is the race, gender, and age group of the person standing?	White, Male, 0-9
	What is the race, gender, and age group of the person with the newspaper?	White, Male, 20-39
	What is the race, gender, and age group of the person on the carpet?	White, Male, 0-9
	What is the race, gender, and age group of the white haired person?	White, Female, More than 60
	What is the race, gender, and age group of the person with the crossed fingers on both hands?	Asian, Female, More than 60
	What is the race, gender, and age group of the person holding a child?	Asian, Female, 20-39

Table 24: Several examples using emotion testing prompt.

Image	Prompt	Generated text
	Emotion of each person	The person on the left appears to be happy. The person in the middle appears to be surprised. The person on the right's emotion is not visible.
	Emotion of the individual wearing the grey shirt	Surprise
	Emotion of each person	The person on the left appears to be expressing “happy”. The person in the center seems to be expressing “anger”. The person on the right appears to be exhibiting “surprise”.
	Emotion of the Black individual	Happy
	Emotion of each person	The person on the left appears to be experiencing ‘fear’, while the person on the right appears to be ‘happy’.
	Emotion of the Middle Eastern individual	Happy
	Emotion of each person	Person on the left: surprise. Person in the middle: happy. Person on the right: fear.
	Emotion of the Asian Eastern individual	Happy
	Emotion of each person	Person in the foreground: happy. Person in the background: disgust

References

- [1] Karkkainen, K. & Joo, J. 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*, 1548–1558 (2021).
- [2] Yuan, H., He, Y., Du, P. & Song, L. Multi-task learning using uncertainty to weigh losses for heterogeneous face attribute estimation. *arXiv preprint arXiv:2403.00561* (2024).
- [3] Vaidya, A., Mai, F. & Ning, Y. Empirical analysis of multi-task learning for reducing identity bias in toxic comment detection. In *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 14, 683–693 (2020).
- [4] Lee, S. H., Plataniotis, K. N. & Ro, Y. M. Intra-class variation reduction using training expression images for sparse representation based facial expression recognition. *IEEE Transactions on Affective Computing* **5**, 340–351 (2014).
- [5] Caruana, R. Multitask learning. *Machine learning* **28**, 41–75 (1997).
- [6] Li, G.-Z., Liu, T.-Y. & Wu, G.-F. Improving generalization ability of neural networks ensemble with multi-task learning. *Journal of Computational Information Systems* **2**, 1235–1239 (2006).
- [7] Han, H., Jain, A. K., Wang, F., Shan, S. & Chen, X. Heterogeneous face attribute estimation: A deep multi-task learning approach. *IEEE transactions on pattern analysis and machine intelligence* **40**, 2597–2609 (2017).
- [8] Mohammad, S. M. Ethics sheet for automatic emotion recognition and sentiment analysis. *Computational Linguistics* **48**, 239–278 (2022).
- [9] Sham, A. H. *et al.* Ethical ai in facial expression analysis: racial bias. *Signal, Image and Video Processing* **17**, 399–406 (2023).
- [10] Leslie, D. Understanding bias in facial recognition technologies. *arXiv preprint arXiv:2010.07023* (2020).

- [11] Li, S. & Deng, W. A deeper look at facial expression dataset bias. *IEEE Transactions on Affective Computing* **13**, 881–893 (2020).
- [12] Schmitz, M., Ahmed, R. & Cao, J. Bias and fairness on multimodal emotion detection algorithms. *arXiv preprint arXiv:2205.08383* (2022).
- [13] Brown, T. *et al.* Language models are few-shot learners. *Advances in neural information processing systems* **33**, 1877–1901 (2020).
- [14] Team, G. *et al.* Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805* (2023).
- [15] Liu, H., Li, C., Wu, Q. & Lee, Y. J. Visual instruction tuning. *Advances in neural information processing systems* **36** (2024).
- [16] Beyer, L. *et al.* Paligemma: A versatile 3b vlm for transfer. *arXiv preprint arXiv:2407.07726* (2024).
- [17] Xiao, B. *et al.* Florence-2: Advancing a unified representation for a variety of vision tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 4818–4829 (2024).
- [18] Kevian, D. *et al.* Capabilities of large language models in control engineering: A benchmark study on gpt-4, claude 3 opus, and gemini 1.0 ultra. *arXiv preprint arXiv:2404.03647* (2024).
- [19] Dalvi, C., Rathod, M., Patil, S., Gite, S. & Kotecha, K. A survey of ai-based facial emotion recognition: Features, ml & dl techniques, age-wise datasets and future directions. *Ieee Access* **9**, 165806–165840 (2021).
- [20] Demszky, D. *et al.* Using large language models in psychology. *Nature Reviews Psychology* **2**, 688–701 (2023).
- [21] Casado, F. E., Lema, D., Iglesias, R., Regueiro, C. V. & Barro, S. Ensemble and continual federated learning for classification tasks. *Machine Learning* **112**, 3413–3453 (2023).
- [22] Minaee, S. *et al.* Large language models: A survey. *arXiv preprint arXiv:2402.06196* (2024).

- [23] Li, C. *et al.* Multimodal foundation models: From specialists to general-purpose assistants. *Foundations and Trends® in Computer Graphics and Vision* **16**, 1–214 (2024).
- [24] Schroff, F., Kalenichenko, D. & Philbin, J. Facenet: A unified embedding for face recognition and clustering. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 815–823 (2015).
- [25] Mustapha, M. F., Mohamad, N. M., Osman, G. & Ab Hamid, S. H. Age group classification using convolutional neural network (cnn). In *Journal of Physics: Conference Series*, vol. 2084, 012028 (IOP Publishing, 2021).
- [26] He, K., Zhang, X., Ren, S. & Sun, J. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, 770–778 (2016).
- [27] AlDahoul, N., Rahwan, T. & Zaki, Y. Ai-generated faces free from racial and gender stereotypes. *arXiv preprint arXiv:2402.01002* (2024).
- [28] Cao, Q., Shen, L., Xie, W., Parkhi, O. M. & Zisserman, A. Vggface2: A dataset for recognising faces across pose and age. In *2018 13th IEEE international conference on automatic face & gesture recognition (FG 2018)*, 67–74 (IEEE, 2018).
- [29] Tan, M. & Le, Q. V. Efficientnet: Rethinking model scaling for convolutional neural networks. *CoRR* **abs/1905.11946** (2019). URL <http://arxiv.org/abs/1905.11946>. **1905.11946**.
- [30] Dosovitskiy, A. *et al.* An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929* (2020).
- [31] Radford, A. *et al.* Learning transferable visual models from natural language supervision. In *International conference on machine learning*, 8748–8763 (PMLR, 2021).
- [32] Mollahosseini, A., Hasani, B. & Mahoor, M. Affectnet: a new database for facial expression, valence, and arousal computation in the wild. *IEEE Trans. Affect. Comput.*(99) 1–1 (2008).
- [33] Ning, M., Salah, A. A. & Ertugrul, I. O. Representation learning and identity adversarial training for facial behavior understanding. *arXiv preprint arXiv:2407.11243* (2024).

- [34] Mao, J. *et al.* Poster++: A simpler and stronger facial expression recognition network. *arXiv preprint arXiv:2301.12149* (2023).
- [35] Savchenko, A. V., Savchenko, L. V. & Makarov, I. Classifying emotions and engagement in online learning based on a single facial expression recognition neural network. *IEEE Transactions on Affective Computing* **13**, 2132–2143 (2022).
- [36] Li, J., Nie, J., Guo, D., Hong, R. & Wang, M. Emotion separation and recognition from a facial expression by generating the poker face with vision transformers. *arXiv preprint arXiv:2207.11081* (2022).
- [37] Abdulwahid, A. A. Classification of ethnicity using efficient cnn models on morph and feret datasets based on face biometrics. *Applied Sciences* **13**, 7288 (2023).
- [38] Mello-Klein, C. Facebook’s ad delivery algorithm is discriminating based on race, gender and age in photos, northeastern researchers find. <https://news.northeastern.edu/2022/10/25/facebook-algorithm-discrimination/> (2022).
- [39] Sunitha, G. *et al.* Intelligent deep learning based ethnicity recognition and classification using facial images. *Image and Vision Computing* **121**, 104404 (2022).
- [40] Ahmed, M. A., Choudhury, R. D. & Kashyap, K. Race estimation with deep networks. *Journal of King Saud University-Computer and Information Sciences* **34**, 4579–4591 (2022).
- [41] Al-Humaidan, N. A. & Prince, M. A classification of arab ethnicity based on face image using deep learning approach. *IEEE Access* **9**, 50755–50766 (2021).
- [42] AlBdairi, A. J. A. *et al.* Face recognition based on deep learning and fpga for ethnicity identification. *Applied Sciences* **12**, 2605 (2022).
- [43] Haseena, S. *et al.* Prediction of the age and gender based on human face images based on deep learning algorithm. *Computational and Mathematical Methods in Medicine* **2022**, 1413597 (2022).

- [44] Fayyaz, M. *et al.* Pedestrian gender classification on imbalanced and small sample datasets using deep and traditional features. *Neural Computing and Applications* **35**, 11937–11968 (2023).
- [45] Sonthi, V. K. *et al.* A deep learning technique for smart gender classification system. In *2023 3rd International Conference on Advance Computing and Innovative Technologies in Engineering (ICACITE)*, 983–987 (IEEE, 2023).
- [46] Yaman, D., Eyiokur, F. I. & Ekenel, H. K. Multimodal soft biometrics: combining ear and face biometrics for age and gender classification. *Multimedia Tools and Applications* 1–19 (2022).
- [47] Tunc, A., Tasdemir, S., Koklu, M. & Cinar, A. C. Age group and gender classification using convolutional neural networks with a fuzzy logic-based filter method for noise reduction. *Journal of Intelligent & Fuzzy Systems* **42**, 491–501 (2022).
- [48] Ciobotaru, A. *et al.* Comparing deep learning and genetic algorithms techniques for age and gender classification. In *2023 3rd International Conference on Electrical, Computer, Communications and Mechatronics Engineering (ICECCME)*, 1–6 (IEEE, 2023).
- [49] Krizhevsky, A., Sutskever, I. & Hinton, G. E. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems* **25** (2012).
- [50] Reddy, A. M., Krishna, V. V., Sumalatha, L. & Obulesh, A. Age classification using motif and statistical features derived on gradient facial images. *Recent Advances in Computer Science and Communications (Formerly: Recent Patents on Computer Science)* **13**, 965–976 (2020).
- [51] Duan, M., Li, K., Yang, C. & Li, K. A hybrid deep learning cnn–elm for age and gender classification. *Neurocomputing* **275**, 448–461 (2018).
- [52] Huang, G.-B., Zhu, Q.-Y. & Siew, C.-K. Extreme learning machine: theory and applications. *Neurocomputing* **70**, 489–501 (2006).
- [53] Chowdary, M. K., Nguyen, T. N. & Hemanth, D. J. Deep learning-based facial emotion recognition for human–computer interaction applications. *Neural Computing and Applications* **35**, 23311–23328 (2023).

- [54] Simonyan, K. & Zisserman, A. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).
- [55] Pandi, C., Adi Narayana Reddy, K., Alladi, R., Chandra Sekhar Reddy, V. & Sumithabhashini, P. Emotion and gender classification using convolution neural networks. In *ICT Systems and Sustainability: Proceedings of ICT4SD 2021, Volume 1*, 563–573 (Springer, 2022).
- [56] Khellat-Kihel, S., Muhammad, J., Sun, Z. & Tistarelli, M. Gender and ethnicity recognition based on visual attention-driven deep architectures. *Journal of Visual Communication and Image Representation* **88**, 103627 (2022).
- [57] Foggia, P., Greco, A., Saggese, A. & Vento, M. Multi-task learning on the edge for effective gender, age, ethnicity and emotion recognition. *Engineering Applications of Artificial Intelligence* **118**, 105651 (2023).
- [58] Abdelhamed, A., Afifi, M. & Go, A. What do you see? enhancing zero-shot image classification with multimodal large language models. *arXiv preprint arXiv:2405.15668* (2024).
- [59] Hello gpt-4o. <https://openai.com/index/hello-gpt-4o/> (2024).
- [60] Devlin, J., Chang, M.-W., Lee, K. & Toutanova, K. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).
- [61] Yang, B. *et al.* Vision transformer-based visual language understanding of the construction process. *Alexandria Engineering Journal* **99**, 242–256 (2024).
- [62] Yildirim, N. *et al.* Multimodal healthcare ai: identifying and designing clinically relevant vision-language applications for radiology. In *Proceedings of the CHI Conference on Human Factors in Computing Systems*, 1–22 (2024).
- [63] Hartsock, I. & Rasool, G. Vision-language models for medical report generation and visual question answering: A review. *arXiv preprint arXiv:2403.02469* (2024).

- [64] Antol, S. *et al.* Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, 2425–2433 (2015).
- [65] Gemini Team, G. Gemini 1.5 technical report. https://storage.googleapis.com/deepmind-media/gemini/gemini_v1_5_report.pdf (2024).
- [66] Introducing gemini 1.5, google’s next-generation ai model. <https://blog.google/technology/ai/google-gemini-next-generation-model-february-2024/#architecture> (2024).
- [67] Gpt-4o: The comprehensive guide and explanation. <https://blog.roboflow.com/gpt-4o-vision-use-cases/> (2024).
- [68] Llava: Large language and vision assistant explained. <https://encord.com/blog/llava-large-language-vision-assistant/> (2024).
- [69] Liu, H. *et al.* Llava-next: Improved reasoning, ocr, and world knowledge (2024). URL <https://llava-vl.github.io/blog/2024-01-30-llava-next/>.
- [70] Dang, P. Multimodal (visual and language) understanding with llava-next. <https://rocm.blogs.amd.com/artificial-intelligence/llava-next/README.html> (2023).
- [71] Zhu, H. *et al.* Harnessing large vision and language models in agriculture: A review. *arXiv preprint arXiv:2407.19679* (2024).
- [72] Thomee, B. *et al.* Yfcc100m: The new data in multimedia research. *Communications of the ACM* **59**, 64–73 (2016).
- [73] Yang Song, Z. Z. Utkface dataset. <https://susanqq.github.io/UTKFace/>.
- [74] AWS. Amazon rekognition. https://docs.aws.amazon.com/rekognition/latest/APIReference/API_AgeRange.html.
- [75] Hugging face. <https://huggingface.co/>.