CHAIN-OF-THOUGHT PROMPTING FOR DEMOGRAPHIC INFERENCE WITH LARGE MULTIMODAL MODELS

Yongsheng Yu[1]   *Jiebo Luo*[1]

[1]Department of Computer Science, University of Rochester
yyu90@ur.rochester.edu; jluo@cs.rochester.edu

ABSTRACT

Conventional demographic inference methods have predominantly operated under the supervision of accurately labeled data, yet struggle to adapt to shifting social landscapes and diverse cultural contexts, leading to narrow specialization and limited accuracy in applications. Recently, the emergence of large multimodal models (LMMs) has shown transformative potential across various research tasks, such as visual comprehension and description. In this study, we explore the application of LMMs to demographic inference and introduce a benchmark for both quantitative and qualitative evaluation. Our findings indicate that LMMs possess advantages in zero-shot learning, interpretability, and handling uncurated 'in-the-wild' inputs, albeit with a propensity for off-target predictions. To enhance LMM performance and achieve comparability with supervised learning baselines, we propose a Chain-of-Thought augmented prompting approach, which effectively mitigates the off-target prediction issue.

Index Terms— demographic inference, Chain-of-Thought prompting, large vision-language models, and multimodal understanding

Fig. 1: Analysis of traditional Supervised Learning (SL) methods and naive LMMs in demographic inference task.

1. INTRODUCTION

Demographic inference [1] involves analyzing population data based on characteristics like age [2, 3], gender [4], and ethnicity [5, 6]. Essential in fields such as sociology, marketing, and public health, it helps identify societal trends, understand consumer behavior, and inform public policies. It is crucial for addressing issues such as aging populations and population migration, significantly impacting societal and economic strategies.

Despite these advancements, current traditional artificial intelligence approaches to demographical inference typically rely on domain-specific, real-world labeled paired data. These methods are often limited in scope, lacking a comprehensive understanding of the Knowledge of Demographic Data, diversity among individuals and groups, cultural backgrounds, and macro socio-economic contexts. As a result, they fail to ac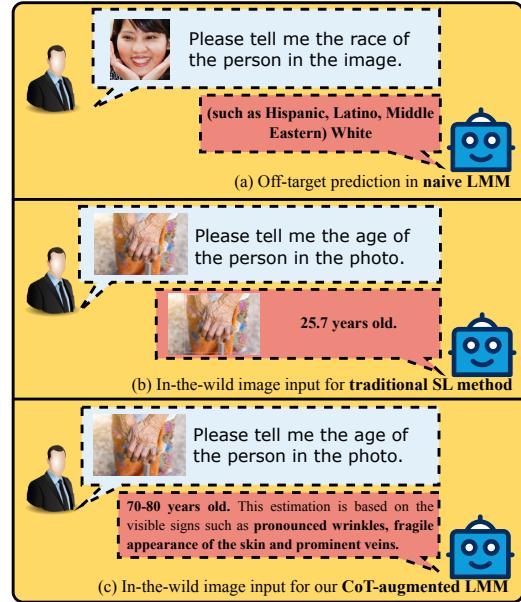curately predict demographic information in in-the-wild data and do not offer an interpretable process or suggestions for demographic inference.

Recently, the emergence of AI foundational models, led by Large Language Models (LLMs) [7], has provided a new paradigm. Characterized by their massive parameter count, training on broad and diverse data, and exceptional versatility, these models can adapt to a wide range of tasks through further training, such as evolving into large multimodal models (LMMs) [8, 9] with the training of visual encoding models. This approach, capable of understanding and processing both image and text inputs, introduces a new paradigm in methodological design across various research fields.

In this study, we propose an integrated demographic inference benchmark and evaluate it on a series of popular open-source LMMs, namely LLaVA [8], MiniGPTv2 [9], Instruct-BLIP [10], and internLM [11]. These LMMs have already demonstrated impressive capabilities in understanding visual content and answering questions with a broad common-sense
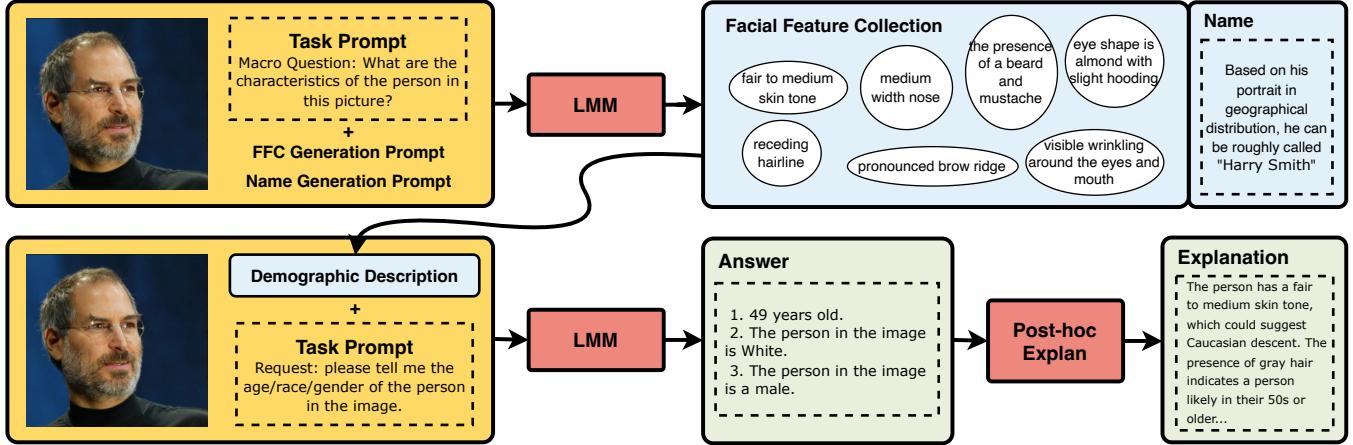
Fig. 2: Conceptual workflow of our Chain-of-Thought prompting approach for demographic inference. The process begins with task prompts guiding LMM to articulate the facial features of the individual in the image, followed by name suggestions. Subsequently, the LMM employs these attributes as demographic descriptions to deduce age, race, and gender, and provides post-hoc explanations for its conclusions.

understanding and high natural language proficiency. Our investigation confirms that LMMs possess three main advantages over traditional demographic inference methods:

1. **Interpretability.** While traditional deep learning approaches yield results conforming to the format of training data labels, LMMs can easily allow the model to explain its predictions through post-hoc questioning.

2. **Zero-shot prediction.** In our study, LMMs do not require any real-world labeled data for downstream task fine-tuning or few-shot data for context-providing instruction and can be directly applied to visual-based demographic inference test datasets.

3. **Proficiency in handling out-of-domain data.** Traditional methods, trained on domain-specific data like cropped standard facial images, struggle with visual inputs like half-body portraits. In contrast, LMMs can accurately predict demographics in in-the-wild images, see Figure 1(b) and (c).

Nevertheless, in-domain test datasets, we have observed a significant gap between the performance of naive LMMs in a zero-shot setting and traditional supervised learning approaches. Additionally, the high degree of response flexibility in the language model of LMMs can lead to off-target predictions (see Fig. 1(a)), a common issue when replacing traditional supervised models with LMMs for predictions. To address this, we also propose a Chain-of-Thought approach, enhancing LMMs' prompting with a two-step intermediary questioning process to obtain demographic feature descriptions. Our main contributions are as follows:

- We are the first, to the best of our knowledge, to incorporate LMMs for demographic inference.

- We propose an integrated benchmark to assess the performance of models in demographic inference and study the effectiveness of LMMs on in-the-wild data.

- We introduce a Chain-of-Thought strategy to improve the performance of LMMs on demographic inference tasks while reducing the rate of off-target predictions.

2. RELATED WORKS

2.1. Demographic Inference

In the evolving landscape of demographic inference, particularly in race, age, and gender prediction, recent advancements have been marked by the integration of deep learning. Gender prediction has similarly benefited from deep learning models [4] applied to ocular and real-world images. In age prediction, the use of discriminant subspace learning [2] extracts aging patterns from facial images, presenting them as distinct manifold structures, thereby enhancing both age estimation and our visual understanding of the aging process. Deep learning systems, especially those trained on time-variant features [4], have shown remarkable results in age prediction from text and images. In this study, our primary focus is on the analysis and interpretation of three key demographic attributes: gender, age, and race.

2.2. Large Multimodal Models

Large Multimodal Models (LMMs) [8, 10] represent a leap in AI technology, synthesizing the nuanced reasoning capabilities of Large Language Models (LLMs) [7] with the perceptual insights of vision-language (VL) models. The transformative aspect of LMMs lies in their ability to perform sophis-
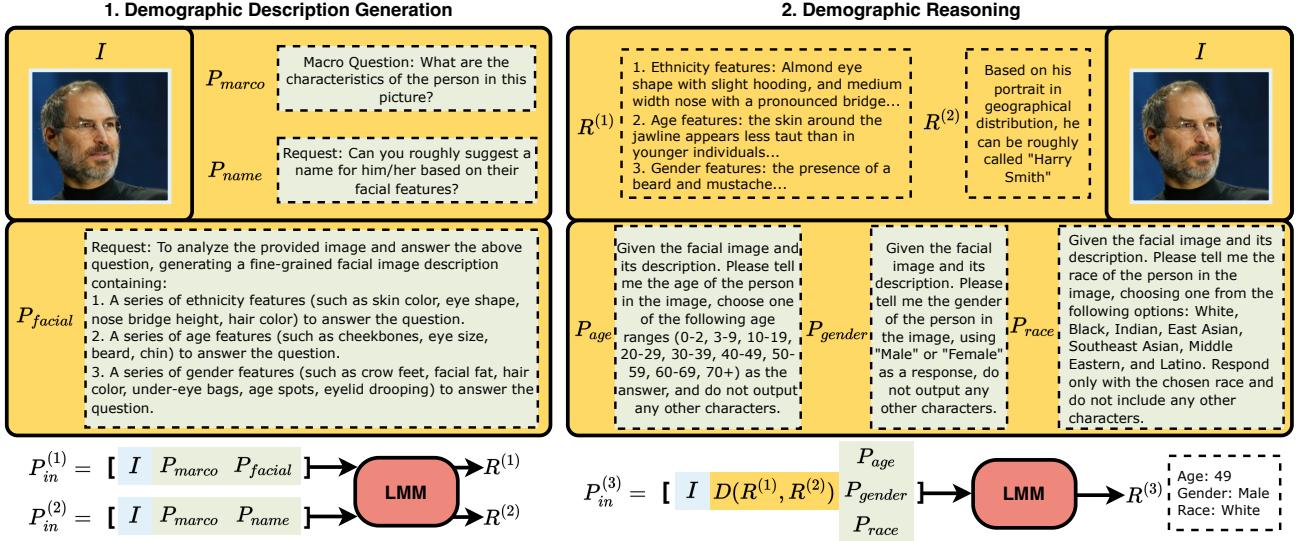
Fig. 3: Full prompt example of the Chain-of-Thought augmented prompting for demographic inference.

ticated tasks that require concurrent understanding and generation of multi-modal data, such as in visual question answering [12] and social media understanding [13]. Our research diverges from traditional LMM applications, pioneering the use of these models for demographic inference. Unlike prior works which primarily leveraged LMMs for tasks like visual question answering or object recognition, our approach explores the untapped potential of LMMs in deciphering demographic information.

2.3. Chain-of-Thought in LLMs

The advent of Chain-of-Thought (CoT) prompting in Large Language Models (LLMs) has emerged as a significant development in AI, enhancing model reasoning capabilities. CoT enables models to process and articulate intermediate steps or 'thoughts' during problem-solving, akin to a human-like reasoning process. This method contrasts with traditional direct answer generation, offering a more transparent and interpretable approach. CoT, along with its extensions like self-consistency, Tree-of-Thought, and Graph-of-Thought, enriches the interaction between users and AI models, particularly in complex tasks requiring detailed explanations or step-by-step reasoning. The implementation of CoT in multimodal contexts further extends its utility, allowing for intricate inference across both textual and visual inputs. This approach demonstrates the evolving sophistication of AI models in mirroring human cognitive processes and their application in diverse problem-solving scenarios.

3. METHODOLOGY

3.1. Problem Formulation

Our task can be conceptualized as a text generation problem under multimodal prompting. The core input to our model is a single-person photo featuring a face, upon which our demographical models infer demographic information such as gender, age, and ethnicity implicitly from visual features, based on textual task instructions. We assess the method's performance using distance metrics for paired data that are either continuous or discrete, aligning with the dataset's label data type. For dataset and experimental settings, please refer to the subsequent sections.

3.2. Large Multimodal Models

Based on text dialogues in large language models $f(\cdot)$ parameterized by θ, LMMs are multimodal models that incorporate both visual and linguistic modalities as inputs. Typically, they receive a collection of images I and a related textual task prompt P, and the large language model, based on the task prompt and visual input, can infer and respond in text form. More specifically, to process visual inputs, language models first map each modality to a shared representational space using pre-trained encoding models, i.e., the visual encoding model embeds I into the textual space, resulting in $\phi(I)$, which is then combined with tokenized language embeddings $\tau(P)$ and fed into the large language model for textual response R.

$$R = f(\phi(I), \tau(P); \theta). \qquad (1)$$

The submodules mentioned above may vary in network architecture, pretraining methods, and parameters across dif-

Table 1: Quantitative experimental results on the UTKFace dataset. Gray row indicates supervised learning baselines.

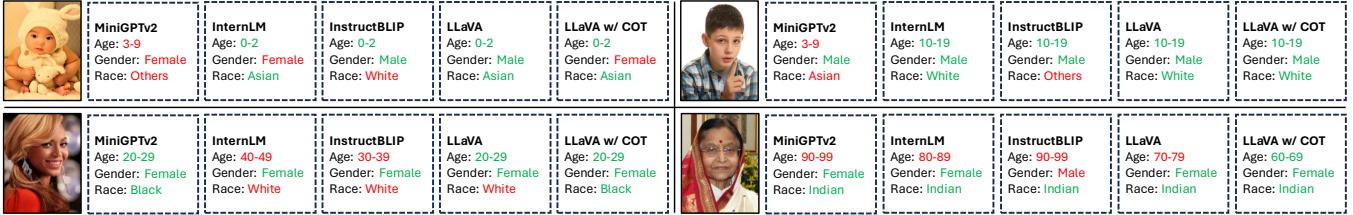|	Age					Gender		Ethnicity		Off-target Rate
	MSE	RMSE	MAE	R^2	MAPE	Accuracy	Kappa	Accuracy	Kappa	
FLAC [6]	-	-	-	-	-	-	-	0.9200	-	0.0%
MWR [14]	-	-	4.37	-	-	-	-	-	-	0.0%
MiniGPTv2 [9]	132.25	11.50	7.28	0.6600	103.36%	0.9540	0.9071	0.5920	0.4656	7.9%
InstructBLIP [10]	241.39	15.54	8.35	0.3794	232.51%	0.8980	0.7914	0.6140	0.4139	20.2%
LLaVA [15]	60.14	7.75	5.13	0.8454	24.03%	0.9620	0.9236	0.8560	0.8005	0.2%
LLaVA w/ COT	55.04	7.42	4.80	0.8585	15.35%	0.9750	0.9496	0.8810	0.8358	0.0%

Fig. 4: Qualitative comparison of naive LMMs and COT-augmented LMMs. Red answers are incorrect, green ones are correct. 'Others' [16] in race categories includes those not White, Black, Asian, or Indian. Zoom in for a better view.

ferent LMMs, such as LLaVA [8] and MiniGPTv2 [9], but generally follow the outlined pipeline.

3.3. Chain-of-thought Augmented Prompting

We find that prompting LMMs with native task instructions does not achieve performance comparable to supervised methods, as shown in Table 1. This is attributed to the excessive versatility of LMMs in handling tasks without being fine-tuned for demographic prediction, where straightforward task instructions fail to fully leverage the prior of LMM in a zero-shot manner.

To maximize the inherent inferential potential of LMMs, we propose enhancing textual prompts with a Chain-of-Thought approach. Our first step is to have LMMs directly interpret the input image, generating a detailed Facial Feature Collection (FFC) P_{facial}, thus avoiding the need for ground-truth caption data. As illustrated in Figure 3, P_{facial} systematically constructs key attributes about age, gender, and ethnicity, e.g., gender-related attributes might include crow's feet, facial fat, hair color, under-eye bags, and age spots.

$$R^{(1)} = f(\phi(I), \tau(P_{\text{marco}}), \tau(P_{\text{facial}}); \theta). \quad (2)$$

Ethnicity is a more challenging category to discern, as predictions rely not only on facial features, especially for an attribute like ethnicity with strong multicultural aspects and common knowledge dependencies (like birth geographical location, and nationality). In the second step, inspired by [1], we consider the person's name as a key feature for discerning ethnicity. Using the naming (captioning) capability of LMMs, we predict the last and first names based on the input image.

We combine the estimated name P_{name} with facial feature collection P_{facial} to form a demographic description D, as seen in Figure 3, which is used to enhance the original textual prompt.

$$R^{(2)} = f(\phi(I), \tau(P_{\text{marco}}), \tau(P_{\text{name}}); \theta). \quad (3)$$

It is noteworthy that in the demographic description, the roles of attributes are not entirely orthogonal; for instance, gender-related attributes might also aid in age prediction. Therefore, in the third step, we integrate the entire demographic description into the prompt for each demographic subcategory prediction as contextual supplementation. Thus, we utilize the generated demographic description as intermediate Chain-of-Thought inference steps, explicitly enhancing the overall input prompt for response generation, as follows:

$$P_{in}^{(3)} = [I, D(R^{(2)}, R^{(2)}), P_{\text{age}|\text{gender}|\text{race}}]. \quad (4)$$

Under Chain-of-Thought enhancement, we modify the previous formula 1, and the final response text R generated by LMMs is as follows:

$$R^{(3)} = f(P_{in}^{(3)}; \phi, \tau, \theta). \quad (5)$$

This leverages the capability LMM of to represent images in high-dimensional spaces and their descriptive power in language modeling. Notably, our enhancement method requires no fine-tuning or few-shot in-context prompting, remaining entirely text-based and zero-shot, and does not require any annotated facial image captions.

Table 2: Comparative performance and off-target prediction mitigation on the FairFace dataset.

(a) Quantitative results on FairFace. Gray row indicates supervised learning baselines.

	Age		Gender		Ethnicity	
	Accuracy	Kappa	Accuracy	Kappa	Accuracy	Kappa
FairFace [5]	0.597	-	0.942	-	0.937	-
MiVOLO [3]	0.611	-	0.957	-	-	-
MiniGPTv2 [9]	0.316	0.175	0.925	0.849	0.472	0.373
InternLM [11]	0.500	0.399	0.955	0.910	0.462	0.351
InstructBLIP [10]	0.291	0.153	0.874	0.744	0.539	0.449
LLaVA [15]	0.499	0.398	0.956	0.912	0.618	0.546
LLaVA w/ COT	0.577	0.490	0.958	0.916	0.692	0.634

(b) Mitigation of off-target prediction on FairFace via COT.

	Off-target Rate	Accuracy
MiniGPTv2	18.4%	0.316
MiniGPTv2 w/ COT	6.8%	0.473
InternLM	0.1%	0.500
InternLM w/ COT	0.0%	0.568
InstructBLIP	17.1%	0.291
InstructBLIP w/ COT	5.5%	0.434
LLaVA	0.0%	0.499
LLaVA w/ COT	0.0%	0.577

Table 3: Quantitative experimental results on the CACD dataset. Gray row indicates supervised learning baselines.

	MSE	RMSE	MAE	R^2	MAPE
CORAL [17]	-	7.48	5.25	-	-
MWR [14]	-	-	4.37	-	-
MiniGPTv2 [9]	92.09	9.60	7.18	0.309	25.24%
InternLM [11]	83.44	9.13	7.39	0.374	25.18%
InstructBLIP [10]	78.44	8.86	6.27	0.412	21.85%
LLaVA [15]	76.19	8.73	6.65	0.428	22.51%
LLaVA w/ COT	66.69	8.17	5.75	0.500	15.99%

4. EXPERIMENTS

This paper integrates a benchmark for evaluating LMMs in demographic inference, incorporating three facial image datasets: CACD [18], FairFace [5], and UTKFace [16].

- **CACD** comprises over 160,000 images of 2,000 celebrities, offering a broad spectrum in terms of age and appearance diversity. The ground truth age labels span from 14 to 54 years.

- **UTKFace** includes more than 20,000 facial images with age, gender, and ethnicity annotations. The age range is extensive, from 0 to 116 years. It encompass a wide array of subjects across different ethnicities and a balance of male and female participants.

- **Fairface** consists of 108,501 images, each labeled with age, gender, and race. It represents seven racial groups: White, Black, Indian, East Asian, Southeast Asian, Middle Eastern, and Latino.

Since there is no need to use these datasets' training sets to fine-tune LMMs for any downstream tasks, we select images from the official test splits of each dataset to serve as our benchmark for evaluating LMMs.

4.1. Models

On our benchmark, we evaluate four popular LMMs: LLaVA [8], InstructBLIP [10], InternLM [11], and MiniGPTv2 [9]. To implement these LMMs, we utilize their official code and pre-trained weights. Specifically, we select the 13B weights version of LLaVA 1.5 [15], the InstructBLIP based on vicuna 13B weights, MiniGPTv2 based on the llama-2 7B [7] language model, and the 7B version of InternLM.

We apply our Chain-of-Thought enhancement to LLaVA. It's noteworthy that commercial models like GPT4V [19] and Gemini, which are not open-sourced and may update versions frequently, are not the focus of this paper regarding performance state-of-the-art (SOTA). Therefore, we only use the aforementioned open-source LMMs, ensuring experimental stability and reproducibility.

4.2. Quantitative analysis

We apply our proposed COT strategy to LLaVA, referred to as "LLaVA w/ COT". For the continuous age annotations in the UTKFace and CACD datasets, we utilize metrics such as MAE, MAPE, and RMSE for evaluation. For other discrete attributes, performance is measured using Accuracy and Kappa. As shown in Figure 2(a) and Figure 1, "LLaVA w/ COT" exhibits top performance on both UTKFace and Fair-Face datasets, particularly in achieving 0.00% off-target rate and high Kappa scores, indicating precise and reliable predictions across age, gender, and ethnicity categories, and significantly outperforming naive LMMs. As shown in Figure 3, "LLaVA w/ COT" variant outperforms LMM baselines in the quantitative analysis on the CACD dataset, exhibiting the lowest Mean Absolute Error (MAE) at 5.75 and the highest R^2 value of 0.500, indicating its strong predictive accuracy and its ability to explain half of the variance in the dataset. Beyond LMMs, our method exhibits comparable performance to traditional supervised learning approaches. In Table 2(a), "LLaVA w/ COT" outperforms state-of-the-art supervised learning methods in gender prediction accuracy.

4.3. Qualitative analysis

In a qualitative comparison, CoT-augmented LMMs show a improvement in demographic prediction accuracy over naive LMMs. As shown in Figure 4, it is evident in the consistent green (correct) labels across the augmented models, indicating precise age, gender, and race identification. The naive LMMs, however, frequently mislabel, particularly in the 'Others' race category, highlighting the difficulty of handling diverse demographic attributes without augmentation. Enhanced models display a higher degree of specificity and sensitivity to visual cues. An interesting observation is that utilizing CoT-augmented prompting for inference may occasionally be counterproductive when dependent on a singe image. For instance, in the top-left case of Figure 4, the CoT-augmented prediction identifies the subject as female, potentially due to the interpretation of clothing details, such as the cute bunny-style outfit, which might be stereotypically associated with female infants. However, the ground truth is male.

4.4. Off-target prediction

Due to the high flexibility of the language model of LMMs, texts generated by them do not necessarily fit the ground truth categories, even with explicit prompting (as shown in Figure 3 with instruction "not to output any other characters"). This off-target phenomenon occasionally enhances the readability of responses or provides some interpretable information, which is acceptable in contexts that do not require quantitative evaluation. However, it also results in ambiguous or irrelevant answers, as seen in Figure 1(a). Therefore, we consider off-target prediction in demographic inference as an issue with LMMs. To fairly assess LMMs, for each demographic inference, if the output is off-target, the inference is repeated until an on-target prediction occurs or N iterations have passed. As shown in Table 2(b), some LMMs still exhibit a high off-target rate despite this procedure. After N repetitions, if the prediction remains off-target, we use the CLIP [20] to encode prediction and measure the similarity between prediction and ground truth categories, selecting the highest similarity as a replacement. We set N to 5 in our experiments. Importantly, we find that the proposed COT method effectively mitigates off-target prediction issues. As depicted in Figure 2(b), MiniGPTv2 with COT sees a reduction in off-target rate from 18.4% to 6.8%, and InstructBLIP with COT drops from 17.1% to 5.5%. Both InternLM and LLaVA achieve a 0.0% off-target rate when combined with COT, underscoring the efficacy of our COT in enhancing model precision.

5. CONCLUSION

Our study presents a comprehensive exploration of LMMs for demographic inference, establishing a benchmark for the field. We demonstrate that LMMs, when enhanced with a Chain-of-Thought prompting strategy, not only provide interpretable, zero-shot predictions that excel in uncurated scenarios but also show promising results in adapting to the diverse and dynamic nature of demographic data. The introduction of this Chain-of-Thought approach significantly reduces off-target predictions, aligning LMM performance closely with that of traditional supervised learning methods, as evidenced by our rigorous quantitative and qualitative evaluations across multiple datasets. Our findings highlight the potential of LMMs to revolutionize demographic inference, offering a flexible and robust alternative to existing AI models. We believe that the integration of such AI models can greatly benefit societal-scale applications, from policy formulation to personalized services, by providing nuanced and accurate demographic analyses.

6. REFERENCES

[1] Neil Yeung, Jonathan Lai, and Jiebo Luo, "Face off: Polarized public opinions on personal face mask usage during the covid-19 pandemic," in *Big Data*. IEEE, 2020.

[2] Yun Fu and Thomas S Huang, "Human age estimation with regression on discriminative aging manifold," *IEEE Transactions on Multimedia*, 2008.

[3] Maksim Kuprashevich and Irina Tolstykh, "Mivolo: Multi-input transformer for age and gender estimation," *arXiv preprint arXiv:2307.04616*, 2023.

[4] Mohamed Ait Abderrahmane, Ibrahim Guelzim, and Abdelkaher Ait Abdelouahad, "Hand image-based human age estimation using a time distributed cnn-gru," in *ICDABI*. IEEE, 2020.

[5] Ni Zhuang, Yan Yan, Si Chen, and Hanzi Wang, "Multi-task learning of cascaded cnn for facial attribute classification," in *ICPR*, 2018.

[6] Ioannis Sarridis, Christos Koutlis, Symeon Papadopoulos, and Christos Diou, "Flac: Fairness-aware representation learning by suppressing attribute-class associations," *arXiv preprint arXiv:2304.14252*, 2023.

[7] Hugo Touvron, Louis Martin, and Kevin Stone et al., "Llama 2: Open foundation and fine-tuned chat models," *CoRR*, vol. abs/2307.09288, 2023.

[8] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee, "Visual instruction tuning," *NeurIPS*, 2023.

[9] Jun Chen, Deyao Li, Zhu Xiaoqian, and Shen Xiang et al., "MINIGPT-V2: Large language model as a unified interface for vision-language multi-task learning," *arXiv preprint arXiv:2310.09478*, 2023.

[10] Wenliang Dai, Junnan Li, and Dongxu Li et al., "Instructblip: Towards general-purpose vision-language models with instruction tuning," 2023.

[11] Pan Zhang, Xiaoyi Dong, and Bin Wang et al., "Internlm-xcomposer: A vision-language large model for advanced text-image comprehension and composition," *CoRR*, vol. abs/2309.15112, 2023.

[12] Daoan Zhang, Junming Yang, Hanjia Lyu, Zijian Jin, Yuan Yao, Mingkai Chen, and Jiebo Luo, "Cocot: Contrastive chain-of-thought prompting for large multimodal models with multiple image inputs," vol. abs/2401.02582, 2024.

[13] Hanjia Lyu, Jinfa Huang, Daoan Zhang, Yongsheng Yu, Xinyi Mou, Jinsheng Pan, Zhengyuan Yang, Zhongyu Wei, and Jiebo Luo, "Gpt-4v(ision) as A social media analysis engine," vol. abs/2311.07547, 2023.

[14] Nyeong-Ho Shin, Seon-Ho Lee, and Chang-Su Kim, "Moving window regression: A novel approach to ordinal regression," in *CVPR*, 2022.

[15] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee, "Improved baselines with visual instruction tuning," *arXiv preprint arXiv:2310.03744*, 2023.

[16] Zhifei Zhang, Yang Song, and Hairong Qi, "Age progression/regression by conditional adversarial autoencoder," in *CVPR*, 2017.

[17] Wenzhi Cao, Vahid Mirjalili, and Sebastian Raschka, "Rank consistent ordinal regression for neural networks with application to age estimation," *PR*, 2020.

[18] Bor-Chun Chen, Chu-Song Chen, and Winston H Hsu, "Cross-age reference coding for age-invariant face recognition and retrieval," in *ECCV*, 2014.

[19] OpenAI, "Gpt-4v(ision) system card," 2023.

[20] Alec Radford, Jong Wook Kim, and Chris et al. Hallacy, "Learning transferable visual models from natural language supervision," in *ICML*, 2021.