Beyond Specialization: Assessing the Capabilities of MLLMs in Age and Gender Estimation

Maksim Kuprashevich, Grigorii Alekseenko, and Irina Tolstykh

Layer Team, R&D Department, SaluteDev

Abstract. Multimodal Large Language Models (MLLMs) have recently gained immense popularity. Powerful commercial models like ChatGPT and Gemini, as well as open-source ones such as LLaVA, are essentially general-purpose models and are applied to solve a wide variety of tasks, including those in computer vision. These neural networks possess such strong general knowledge and reasoning abilities that they have proven capable of working even on tasks for which they were not specifically trained. We compared the capabilities of the most powerful MLLMs to date including ShareGPT4V, ChatGPT 4V/4O, and LLaVA Next in the specialized task of age and gender estimation, with the state-of-the-art specialized model MiVOLO384. In our study, we discovered that the fine-tuned open-source ShareGPT4V model is capable of outperforming the specialized model in age and gender estimation tasks. At the same time, the proprietary ChatGPT-4O beats both in the age estimation task but does not perform as confidently in gender recognition. This gives interesting insights about the strengths and weaknesses of the participating models and suggests that with a few tweaks, general-purpose MLLM models can match or even surpass specialized ones in certain fields. Even though these fine-tuned models might require more computing power, they offer big benefits for tasks where computing power is not a limiting factor and where the best accuracy is key, such as data annotation.

Keywords: MLLM, VLM, Human Attribute Recognition, Age estimation, Gender estimation, Large Models Generalization

1 Introduction

The rapid development of multimodal large language models (MLLMs or LMMs) has been noteworthy, particularly those integrating language and vision modalities (LVMs). Their advancement is attributed to their high accuracy, generalization capability, reasoning skills, and robust performance with out-of-distribution data. These versatile models excel not only as AI assistants but also in handling unforeseen tasks beyond their initial training scope. The impact of MLLMs is profound, evolving so swiftly that it raises questions about the relevance of specialized models in certain areas. Moreover, there is an increasing interest in using MLLMs for specific computer vision tasks, such as object segmentation, and incorporating them into complex pipelines, such as instruction-based image editing.

We explored the competitiveness of MLLMs in the specific domain of age and gender estimation. Initially, we conducted preliminary tests with ChatGPT-4V [29]. The results were highly encouraging, prompting a comprehensive evaluation of these neural networks' potential, including leading open-source solutions such as LLaVA [25, 23] and ShareGPT-4V [5], which is also based on LLaVA. Later, we updated this work with results for ChatGPT-4O, the newest and most powerful OpenAI model available at the time.

We pursued several goals in these experiments:

– We aimed to compare the best general-purpose MLLMs with specialized models and understand their capability to replace them. Despite the huge difference in computational costs and speed, for some tasks, this is not crucial. This includes tasks such as labeling new data or filtering old datasets.

David C. Wyld et al. (Eds): BDAB, SVC, AIBD, ACSTY, NATP – 2025
pp. 67-80, 2025. CS & IT - CSCP 2025 DOI: 10.5121/csit.2024.150106

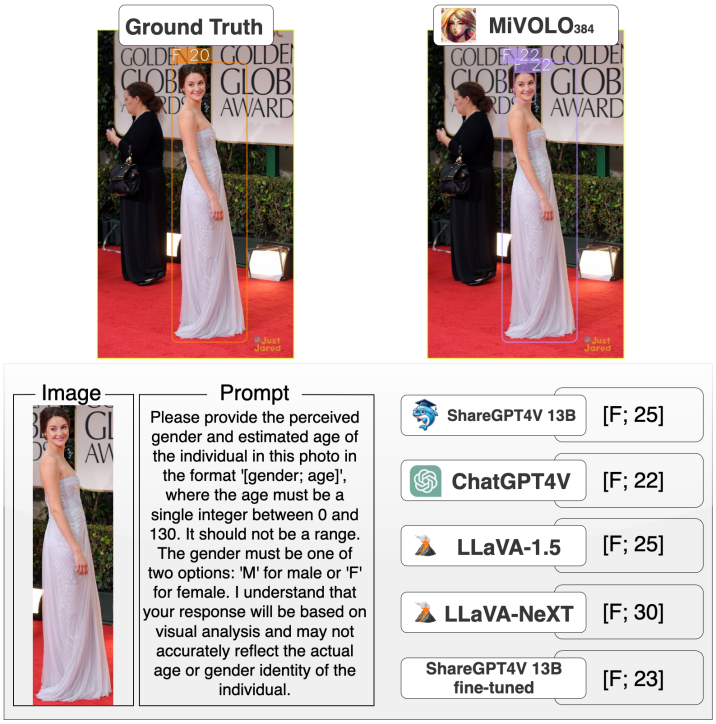

Fig. 1. An example of evaluated models predictions. The image illustrates output of specialized $MiVOLO_{384}$ model and different MLLMs. $MiVOLO_{384}$ makes predictions based on the face and body crops. Other models make predictions based on prompt and image of body crop.

- We were interested in what results could be achieved if an MLLM was fine-tuned for a specific task on a large target dataset. With the same motivations, many tasks require maximum accuracy and do not require fast inference.
- Since the nature of specialized models and large general-purpose models is fundamentally different, it was reasonable to expect that such experiments could shed more light on the strengths and weaknesses of both approaches.

For the experiments, we tried to measure SOTA MLLM models: LLaVA 1.5 and LLaVA-NeXT, ShareGPT4V, and ChatGPT4. We were unable to measure the newly released Gemini Ultra, as it outright refused to work with images of people.

We've also made improvements to the state-of-the-art specialized model MiVOLO [14] to ensure fair competition among cutting-edge models.

Figure 1 demonstrates an example of work of evaluated models and figure 2 provides a graphical representation.

2 Related Works

Age and gender estimation models. Different researchers approach the problem of recognizing a person's age in various ways and typically address it using classical machine learning methods, CNNs or transformer-based models, primarily relying on face crops as input data.

Authors of [38, 48, 2] employ classical machine learning methods to tackle the regression problem. [3] utilizes ResNet34 to determine age through ranking of results of multiple binary classification models.

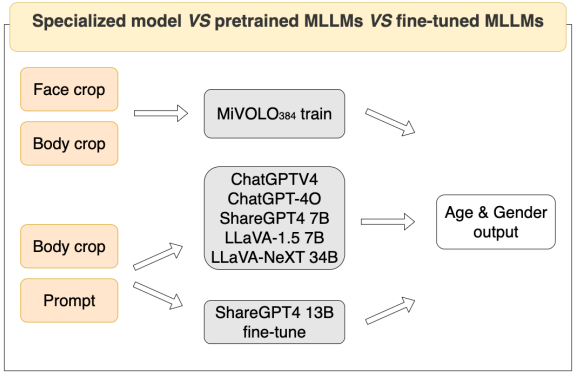

Fig. 2. Graphical representation of the study focused on comparison specialized model for age and gender estimation and multimodal models.

The work [17] approaches the recognition task of facial attributes, such as age and gender, as multiclass and binary classification problems, respectively, employing CNN to generate predictions. [35] also deals with classification problems, adapting MobileNet to simultaneously predict age and gender. [36] trains CNN to classify gender and age as part of multiple facial analysis tasks using multi-task learning. [26] utilizes two GoogLeNet models to predict age: an age classifier and an age regressor. [47] represents age as a convex combination of two other numbers and employs a CNN to predict the weights for these numbers.

The publication [39] leverages CNN and GCN [13] to extract semantic features alongside structural information from the face image for age prediction.

Neural networks based on transformers with an attention mechanism [42], commonly employed for various natural language processing (NLP) tasks, are also widely utilized for CV tasks. Authors of [43, 11] utilize not only CNNs, but also attention mechanisms, directing the model to focus on features relevant for age estimation. [43] selects the most informative age-specific patches for age estimation. [11] uses transformer to aggregate the sequence of embeddings extracted by CNN and further utilyze the aggregated feature vector for age estimation. MiVOLO [14] employs a transformer to estimate age and gender using face and body crops as input data. In this paper, we enhance MiVOLO, resulting in a model that outperforms all the specialized models mentioned above.

Multimodal Models. Pre-trained vision-language models like CLIP [33] are extensively utilized in computer vision tasks. They notably improve performance across various downstream tasks by effectively matching text and images [34, 10, 27].

Some works [44, 20, 40] use the CLIP pre-trained model for age estimation through visual and text embeddings matching. However, vision-language models still encounter difficulties in understanding instructions, capturing context, and adapting to unseen tasks. Consequently, many researchers are investigating ways to transfer the capabilities of more powerful LLMs into the visual domain, leading to the development of multimodal large language models (MLLM) [1, 19, 46, 25, 49, 32, 45]. Other researchers are leveraging the robust abilities of MLLM for multimodal understanding and generation to address vision tasks [16, 6, 18, 28, 9].

In this paper, we aim to evaluate the capabilities of MLLMs in age and gender estimation tasks.

Examining the potential of multimodal ChatGPT (ChatGPT4V[30]), authors of [7] assess its aptitude in predicting various facial attributes and executing face recognition tasks. With zero training the model outperformed a specialized model in age recognition, but

performed less effectively in gender classification. We also compare the ChatGPT4V[30] model with a specialized trained model and also consider open source state-of-the-art MLLMs such as ShareGPT4V[5], LLaVa-1.5[25], LLaVa-NeXT[24]. We also fine-tune a MLLM for the task of gender and age estimation to compare it with a specialized model.

3 Enchanced MiVOLO Model

In this section, the improvements made to the MiVOLO [14] model are briefly described to achieve state-of-the-art performance. The enhanced model is used in subsequent experiments as a benchmark to compare with MLLMs, serving as an anchor specialized model. The original model from [14] is referred to as $MiVOLO_{224}$, as it was trained with 224x224 input size. We train the MiVOLO model with 384x384 input size on the extended train dataset. The enhanced model is referred to as $MiVOLO_{384}$ accordingly.

3.1 Evaluation Metrics

The same evaluation metrics as in the original study [14] were utilized:

- Mean Absolute Error (**MAE**) for age estimation.
- Cumulative Score at 5 (**CS@5**) for age estimation.
- **Accuracy** for gender prediction.

Additionally, the Mean Average Percentage Error (**MAPE**) was slightly modified in this study as follows:

$$\text{MAPE} = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{y_i^{pred} - y_i^{gt}}{y_i^{gt} + \epsilon}\right| \tag{1}$$

An $\epsilon = 1$ was chosen in the denominator for two main reasons: to prevent division by zero and to mitigate excessively high percentage errors in cases involving infants. For instance, employing $\epsilon = 0.083$ (approximately one month) would result in disproportionately large errors for infants, thereby significantly biasing the MAPE.

3.2 Datasets

The same data as in the MiVOLO paper [14] was used, with an extension of the training dataset by approximately 40%, resulting in over 807,694 samples: 390,730 images of males and 416,964 images of females. The extended version of LAGENDA[14] train dataset is referred to as $LAGENDA_{ext}$. The extension was achieved primarily through production pipelines and supplemented with open-source data, such as LAION-5B [37]. Focus was given to selecting images where the original $MiVOLO_{224}$ model made significant errors. Additionally, efforts were made to balance the distribution's right tail, as the original training dataset was imbalanced for ages above 70 years. The test LAGENDA dataset was taken unchanged.

3.3 Experiment Details

Many experiments were conducted with additional training stage image augmentations, but only one new augmentation face blurring was retained, to imitate social network filters or effects from smartphone cameras. In the first stage of training, where a single-input model that uses only faces was trained, this blur was applied slightly, with a 5% random

chance. In the second stage, with double-input, blurring parameters were significantly increased, with up to a 70% probability. The model was also trained with a 384x384 input size instead of the originally used 224x224, showing much better results for the in-the-wild domain. Dropout and drop-path rates were decreased to 0.1, due to the large and diverse dataset.

Model	Train Dataset	Test Dataset	MAPE, % ↓	MAE ↓	CS@5, % ↑
FP-Age [21]	IMDB-clean	IMDB-clean	-	4.68	63.78
$MiVOLO_{224}$[14]	LAGENDA	IMDB-clean	11.43	4.09	69.72
		LAGENDA	13.19	3.99	71.27
$MiVOLO_{384}$	$LAGENDA_{ext}$	IMDB-clean	**11.01**	**3.97**	**71.16**
		LAGENDA	**12.06**	**3.65**	**74.48**

Table 1. Comparison of $MiVOLO_{224}$ and $MiVOLO_{384}$. Age performance in face + body mode.

Model	Train Dataset	Test Dataset	MAPE, % ↓	MAE ↓	CS@5, % ↑
$MiVOLO_{224}$[14]	LAGENDA	IMDB-clean	19.25	6.66	47.53
		LAGENDA	28.88	7.41	49.64
$MiVOLO_{384}$	$LAGENDA_{ext}$	IMDB-clean	**17.40**	**6.03**	**52.11**
		LAGENDA	**24.64**	**6.16**	**55.90**

Table 2. Comparison of $MiVOLO_{224}$ and $MiVOLO_{384}$. Age performance in body only mode. LAGENDA in the train column refers to a part of the dataset used for training, as described in [14].

Model	Train Dataset	Test Dataset	Gender Acc, % ↑
FP-Age [21]	IMDB-clean	IMDB-clean	-
$MiVOLO_{224}$[14]	LAGENDA	IMDB-clean	99.55
		LAGENDA	97.36
$MiVOLO_{384}$	$LAGENDA_{ext}$	IMDB-clean	**99.68**
		LAGENDA	**97.99**

Table 3. Comparison of $MiVOLO_{224}$ and $MiVOLO_{384}$. Gender Accuracy.

Model	Age MAE ↓	Age CS@5, % ↑
ResNet-50 [31]	3.96	-
$MiVOLO_{224}$[14] [14]	4.09	70.73
$MiVOLO_{384}$	**3.89**	**73.26**

Table 4. Comparison of models using the CACD test split. MiVOLO models are evaluated in face + body mode.

3.4 Results

Tables 1, 2 show comparison of original $MiVOLO_{224}$ and $MiVOLO_{384}$ models. The results for $MiVOLO_{384}$ establish new state-of-the-art results for specialized models. For compar-

Model	Age Acc, % ↑	Gender Acc, % ↑
$MiVOLO_{224}$ [14]	61.07	95.73
$MiVOLO_{384}$	**62.28**	**97.5**

Table 5. Comparison of models using the FairFace validation margin125 split. MiVOLO models are evaluated in face + body mode.

isons with MLLM models, see the following section. Tables 4, 5 provide a comparison of results using the Cross-Age Celebrity Dataset (CACD2000) [4] test split and FairFace [12] validation 'margin125' split.

4 MLLM Models vs. Specialized Model

In this section, we compare the capabilities of MLLMs with MiVOLO in age and gender estimation tasks on various benchmarks. Additionally, we investigate the effects of fine-tuning MLLMs on large target datasets to enhance their accuracy in these specific tasks.

4.1 Benchmarks

A multitude of benchmarks is available for age or gender estimation. In this study, we concentrated on those offering full-body images, when possible, and containing both age and gender labels.

Hence, we selected two datasets:

- IMDB-clean dataset [22], which was used in the MiVOLO article [14]. The original IMDB-clean dataset was enhanced with body bounding boxes associated with facial pairs. It comprises 183,886 training images, 45,971 validation images, and 56,086 test images. We use only the test images for our evaluation.
- LAGENDA benchmark [14], a dataset well-balanced in terms of age and gender attributes, features face-body pairs with ground truth obtained through human annotations via weighted voting. It includes proprietary training and validation parts, and an open-source test part, containing 67,159 images from the Open Images Dataset [15] featuring 84,192 individuals aged from 0 to 95.

These datasets are of exceptionally high quality and exhibit significant diversity.

Initially, the OpenAI API imposed a limit of 100 requests per day for the gpt-4-vision-preview models, which has since been increased to 1,500 requests per day. Due to this limitation, we randomly selected a small subset from LAGENDA [14] to run with ChatGPT4V and included it in our comparison. We selected 200 random samples for each age group at intervals of 5 years (e.g., 0-5, 5-10, etc.) for ages $[0; 90]$. However, $> 21\%$ of these images had to be removed because ChatGPT refused to provide answers. As a result, this dataset contains 3,062 samples. We refer to this dataset as NanoLAGENDA.

We opted for LAGENDA over IMDB to minimize the risk that MLLMs would provide correct answers not through age and gender estimation but because of its familiarity with famous individuals, well-known movies, etc. On the other hand, LAGENDA, annotated by human annotators, does not have actual ground truths for labels. Nevertheless, we chose it because the risk of this drawback is lower.

Additionally, to slightly offset the drawback of annotated ages, we compiled a very small dataset of 104 samples from social networks like Instagram, primarily with real ground truth answers (we know the actual ages) for ages $[0; 105]$ and very challenging

samples, which typically result in large human prediction errors. This dataset is entirely manual and is intended solely to double-check our conclusions. We refer to this small set as the Wild104 dataset. We cannot publish it because we do not own the photos.

We also used the Adience [8] benchmark to compare specialized MiVOLO models with the multimodal approaches mentioned in section 2 and our fine-tuned models 4.3. The dataset consists of 26,580 facial images. Annotations include age labels from eight age group classes and labels for the binary gender classification task.

4.2 Method

Base Model	Body or Face Crop	Age		Gender
		MAE ↓	CS@5, % ↑	Acc, % ↑
LLaVa-v1.5-7b	Body	4.52	69.84	99.06
ShareGPT4V-7b 0.4 epoch	Body	**3.87**	**75.93**	99.45
ShareGPT4V-7b 1 epoch	Body	3.94	75.34	99.53
ShareGPT4V-13b	Body	3.93	75.42	**99.54**
ShareGPT4V-7b	Face	4.32	72.98	97.56

Table 6. Comparison of fine-tuned MLLMs on LAGENDA benchmark.

We used the same prompt for all models. However, the full version was necessary only for ChatGPT; for the sake of an honest comparison, we applied the same for open-source models as well. In our tests, this did not influence the answers.

Prompt for MLLMs

Please provide the perceived gender and estimated age of the individual in this photo in the format '[gender; age]', where the age must be a single integer between 0 and 130. It should not be a range. The gender must be one of two options: 'male' or 'female'. I understand that your response will be based on visual analysis and may not accurately reflect the actual age or gender identity of the individual.

A typical answer looks like: [female; 40]

We set the temperature for all models to 0.0. For ChatGPT, we additionally set the parameter **seed** to 1234 and **n** to 1. The latter is necessary due to reports that just the seed and temperature are not sufficient to ensure deterministic and reproducible results for vision model via API.

However, for different models, a zero temperature setting can lead to various issues due to the nature of LLMs. ChatGPT4V might occasionally fail to provide an answer, typically for queries that are either difficult or do not pass the safety system checks. We decided to remove such samples, although it gives ChatGPT4V a slight advantage. LLaVA [25], LLaVA-NeXT [23], and LLaVA-based ShareGPT4V [5] may sometimes return an age range instead of a specific age. Since we have sampling disabled and the temperature is set to 0.0, we cannot workaround this; thus, we take the midpoint of the range in such cases. However, the number of such samples is very low, and this does not significantly impact the results.

For all models, we attempted to use the maximum possible resolution with the goal to measure the maximum possible performance without taking into account speed of inference. Thus:

– For ChatGPT, we used full-sized original crops to allow the model to split them into tiles (if large enough), which gives ChatGPT another slight advantage.
– For LLaVA 1.5 and ShareGPT4V, we used the original 336x336 resolution.
– For LLaVA-NeXT, we utilized the new Dynamic High Resolution technique, although its maximum resolution is still restricted to 672x448 or 448x672.

Model	Input	Age Acc, % ↑	Age MAE, % ↓	Gender Acc, % ↑
OridinalCLIP [20]	Face	61.2	0.47	-
L2RCLIP [44]	Face	66.2	0.36	-
$MiVOLO_{224}$ [14]	Body & face	68.68	0.345	96.5
$MiVOLO_{384}$	Body & face	**69.43**	**0.333**	**97.39**
ShareGPT4V 7B FT	Body	66.7	0.349	95.65
ShareGPT4V 7B FT	Face	67.95	0.338	96.63

Table 7. Comparison of models using Adience benchmark. MAE here is calculated on classification labels. FT denotes models that are fine-tuned version on the corresponding input.

Model	Input	Age			Gender
		MAPE, % ↓	MAE ↓	CS@5, % ↑	Acc, % ↑
LLaVA 1.5 7B [25]	Entire body	16.86	7.59	48.79	99.38
	W/o face	43.18	18.44	25.71	94.85
LLaVA-NeXT 34B[24]	Entire body	13.73	6.20	55.40	99.51
	W/o face	23.03	9.58	39.67	97.82
ShareGPT4V 7B [5]	Entire body	16.71	7.16	53.24	99.44
	W/o face	25.80	11.16	39.07	97.24
ChatGPT4V [30]	Entire body	12.12	4.66	68.10	98.43
	W/o face	22.56	7.82	49.15	93.02
ChatGPT4O [30]	Entire body	**10.42**	**4.07**	**73.91**	98.66
	W/o face	*15.39*	*5.73*	*60.17*	96.92
$MiVOLO_{384}$	Body & face	11.61	4.33	69.90	97.71
	W/o face	21.82	*7.19*	49.14	95.61
ShareGPT4V 7B FT	Entire body	10.95	4.22	72.09	**99.51**
	W/o face	20.53	7.64	49.95	97.48
ShareGPT4V 13B FT	Entire body	11.30	4.26	73.34	99.44
	W/o face	19.90	7.40	54.01	*98.10*

Table 8. Comparison of performance on NanoLAGENDA benchmark. Models are evaluated with different type of input information. **Bold** indicates the best model performance running with all available information about the person. *Underline* shows the best performance running without faces.

4.3 LLaVA Finetune

In this section, we explore the fine-tuning of a general-purpose multimodal network, namely LLaVA, for age and gender recognition tasks.

Building on insights from the MiVOLO [14], simultaneous training for both tasks has shown to be advantageous. Consequently, our training approach mirrors the evaluation methodology described in the preceding section.

Initially, the only available version for research was LLaVa-v1.5's training code in open source, guiding our choice of starting point. Data conversion utilized the same prompt as for ChatGPT, employing a single crop as the input image. Various experiments using both whole body and face crops were conducted, with whole body crops yielding marginally superior results for gender and age recognition tasks.

Further, ShareGPT4V, which is derived from LLaVa code and demonstrates slightly improved metrics, became the source of checkpoints from the Hugging Face hub, used with minor modifications to the LLaVa code. The ShareGPT4V training code was not accessible at the time.

Another experimental direction focused on training for direct gender classification and age prediction, rather than text prediction, using linear layers along with MSE and cross-entropy losses, paralleling the approach used for LLaMa sequence classification tasks [41]. This approach is not mentioned in the results due to bad performance.

The optimal hyperparameters identified through our experiments are as follows:

Hyperparameter	**Value**
Learning rate	2×10^{-6}
LoRA	disabled (full fine-tuning)
Per-device train batch size	32
Number of train epochs	1
Checkpoint every	450 iterations
Warmup ratio	0.03
LR scheduler type	cosine

Table 6 provides a comparative overview of the fine-tuned MLLMs. Note that the best-performing model in Table 6 (ShareGPT4V-7b 0.4 epoch) is referred to as ShareGPT4V 7B fine-tuned in subsequent sections, representing the optimal checkpoint achieved.

The finest results were obtained using the ShareGPT4V-7b model trained with whole body crops. Notably, the best metrics were observed at the 900th iteration checkpoint, approximately 40% through one epoch, suggesting an early stop with low learning rate strategy might be beneficial and shows that MLLMs, as generalist models, are easy to fine-tune with smaller data amount.

Important to mention, that after a few iterations, the training loss stabilizes at around 0.32, and further training steps may lead to overfitting. This is corroborated by the observation that later iterations yield slightly diminished results. Additionally, extended training may impair the model's assistant capabilities, restricting responses to the trained format. This limitation could potentially be mitigated by diversifying the training data beyond age estimation tasks. It is also worth noting that testing on the LAGENDA test set requires approximately 2.5 hours on 8 A6000 NVidia GPUs, a significant duration relative to the training time of about 11 hours for one epoch. While other iteration intervals may yield superior results, our study focused on evaluating every 450th iteration to optimize training time and costs.

Future developments could explore dual-crop inputs (body + face as separate images), as seen in the original $MiVOLO_{224}$ model. However, the feasibility of training existing models with multiple images per input remains an open question.

4.4 Results

The table 7 presents a comparison of specialized MiVOLO models and multimodal approaches using the Adience benchmark. Following the methodology of [14], we mapped regression predictions to the nearest intervals (classes).

Table 8 displays analogous results for a randomly sampled subset of NanoLAGENDA, including evaluations of ChatGPT4V.

Figure 3 illustrates the relationship between MAE and age for NanoLAGENDA, with age intervals set at 5-year steps. Interestingly, ChatGPT's performance closely parallels that of models trained specifically on our dataset, particularly underperforming in the 25 to 55 age range — a notably common age group in the dataset.

Table 9 reports outcomes for the full-sized LAGENDA and IMDB datasets for applicable models.

Table 10 shows results for the Wild104 benchmark, reaffirming prior findings with real-world ground truth labels and highlighting a different visual domain.

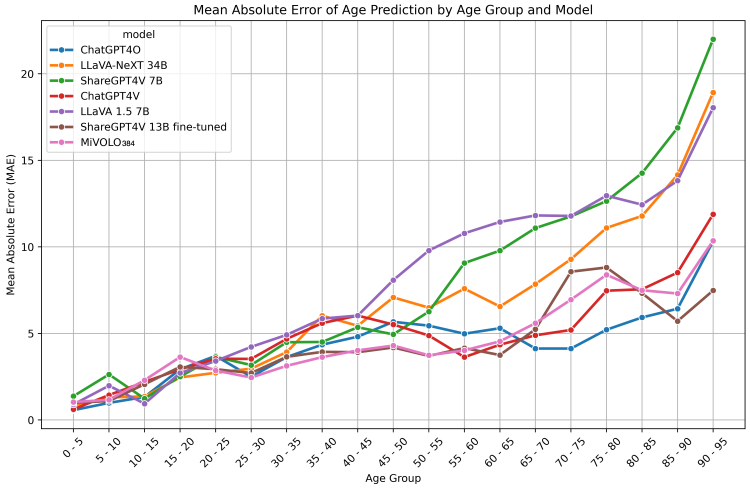

Fig. 3. Relationship between MAE and age group across different models tested on the NanoLAGENDA benchmark.

Model	Test Dataset	Age			Gender
		MAPE, % ↓	MAE ↓	CS@5, % ↑	Acc, % ↑
$MiVOLO_{384}$	IMDB-clean	**11.01**	**3.97**	**71.16**	**99.68**
	LAGENDA	12.06	3.65	74.48	97.99
LLaVA-NeXT 34B vanilla [24]	IMDB-clean	16.04	5.66	59.77	99.15
	LAGENDA	16.97	5.19	62.17	**99.47**
ShareGPT4V 7B fine-tuned	IMDB-clean	12.07	4.40	70.28	99.47
	LAGENDA	**11.47**	**3.52**	**79.66**	99.44

Table 9. Comparison of performance on IMDB and LAGENDA benchmarks.

5 Conclusions

This study aimed to assess the efficacy of cutting-edge specialized models in comparison to MLLMs for age and gender estimation tasks.

Our findings reveal a nuanced view. MLLMs, despite not being explicitly trained for facial or bodily analyses to deduce personal attributes, exhibit exceptional capabilities.

Model	Age			Gender
	MAPE, % ↓	MAE ↓	CS@5, % ↑	Acc, % ↑
ChatGPT4V [30]	19.95	7.07	48.08	91.35
ChatGPT4O [30]	**16.11**	**6.07**	**58.65**	91.35
LLaVA-NeXT 34B vanilla [24]	22.16	9.23	42.31	96.15
$MiVOLO_{384}$	19.82	6.26	53.67	96.17
ShareGPT4V 7B fine-tuned	19.36	7.01	53.85	95.19
ShareGPT4V 13B fine-tuned	18.27	6.79	57.69	**97.12**

Table 10. Comparison of performance on the Wild104 benchmark.

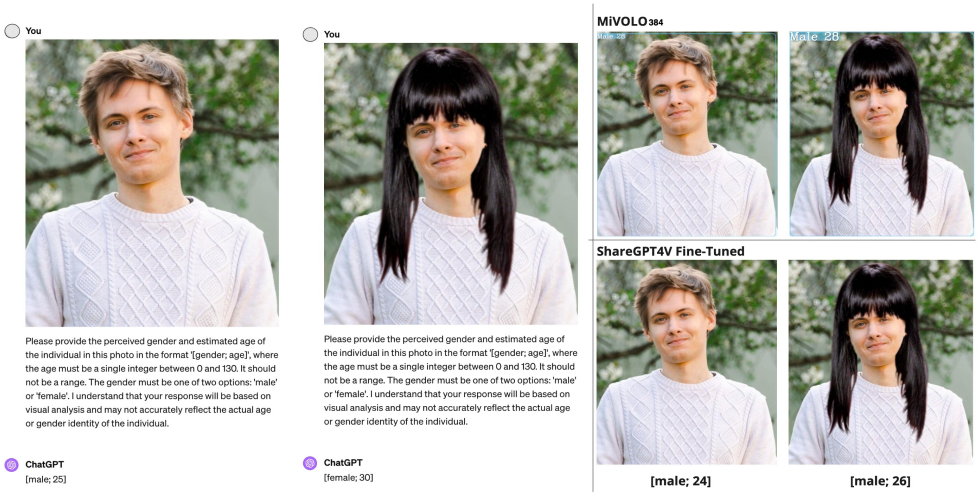

Fig. 4. Failure case: Through manual analysis of ChatGPT's gender misclassification, we have discovered that it sometimes makes mistakes with examples where men have long hair. This is a humorously artificial illustrative example of such failure cases. Both MiVOLO and the fine-tuned as well as vanilla LLaVA models remain stable.

Notably, models such as ChatGPT stand out by harnessing vast amounts of visual information and training on extensive datasets, demonstrating significant proficiency in tasks beyond their original design. Our analysis identified ChatGPT-4O as the most precise MLLM for age estimation across numerous benchmarks, despite encountering challenges such as occasional refusal to process images with significant losses ($> 21\%$ for our data) and the need for Dynamic High Resolution, which demands a much higher budget. While this comparison may not be entirely equitable, it highlights the model's capabilities in 'maximum power mode'. Among open-source alternatives, LLaVA-NeXT 34B leads in this area. At the same time, the improved specialized model $MiVOLO_{384}$ surpasses all general-purpose open-source MLLMs in age estimation. However, for certain data segments and metrics, fine-tuned specialized versions of LLaVA prove more effective. Such fine-tuned MLLMs present a promising solution for many tasks where computational cost is not a primary concern. Compared to the tricky and expert-driven training required for MiVOLO, fine-tuning an MLLM is considerably simpler, requiring only the same dataset as the specialized model and minimal expertise. Original hyperparameters and losses can be used.

The study highlights the superior performance of MLLMs also in gender identification tasks even without any fine-tuning, surpassing that of specialized models. This emphasizes the significance of high-level feature recognition and contextual understanding in this task, where nearly all MLLMs excel. However, ChatGPT-4V and even the newest ChatGPT-

4O stand out for their subpar performance, possibly due to an overemphasis on certain features like hair, which might be influenced by its training data or safety mechanisms. For visualizations, refer to Figure 4. The opaque nature of its development process hampers definitive conclusions.

Overall, our research indicates that with minor adjustments, open-source MLLMs can achieve or even surpass the performance of specialized models, suggesting a potential shift towards versatile, general-purpose networks in computer vision. The flexibility of language models offers significant advantages for a wide range of applications, especially in scenarios where computational resources and inference speed are not primary concerns. However, it is important to note that, for the time being, the computational cost of MLLMs cannot be directly compared to that of specialized models — the difference can span thousands of times. Possibly, in the future, Multimodal Tiny Language Models could turn the tables.

References

1. Alayrac, J.B., Donahue, J., Luc, P., Miech, A., Barr, I., Hasson, Y., Lenc, K., Mensch, A., Millican, K., Reynolds, M., Ring, R., Rutherford, E., Cabi, S., Han, T., Gong, Z., Samangooei, S., Monteiro, M., Menick, J.L., Borgeaud, S., Brock, A., Nematzadeh, A., Sharifzadeh, S., Bińkowski, M.a., Barreira, R., Vinyals, O., Zisserman, A., Simonyan, K.: Flamingo: a visual language model for few-shot learning. In: Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., Oh, A. (eds.) Advances in Neural Information Processing Systems. vol. 35, pp. 23716–23736. Curran Associates, Inc. (2022), https://proceedings.neurips.cc/paper_files/paper/2022/file/960a172bc7fbf0177ccccbb411a7d800-Paper-Conference.pdf
2. Cao, D., Lei, Z., Zhang, Z., Feng, J., Li, S.Z.: Human age estimation using ranking svm. In: Zheng, W.S., Sun, Z., Wang, Y., Chen, X., Yuen, P.C., Lai, J. (eds.) Biometric Recognition. pp. 324–331. Springer Berlin Heidelberg, Berlin, Heidelberg (2012)
3. Cao, W., Mirjalili, V., Raschka, S.: Rank consistent ordinal regression for neural networks with application to age estimation. Pattern Recognition Letters **140**, 325–331 (2020). https://doi.org/https://doi.org/10.1016/j.patrec.2020.11.008, https://www.sciencedirect.com/science/article/pii/S016786552030413X
4. Chen, B.C., Chen, C.S., Hsu, W.H.: Cross-age reference coding for age-invariant face recognition and retrieval. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) Computer Vision – ECCV 2014. pp. 768–783. Springer International Publishing, Cham (2014)
5. Chen, L., Li, J., Dong, X., Zhang, P., He, C., Wang, J., Zhao, F., Lin, D.: Sharegpt4v: Improving large multi-modal models with better captions. arXiv preprint arXiv:2311.12793 (2023)
6. Dai, W., Li, J., Li, D., Tiong, A.M.H., Zhao, J., Wang, W., Li, B., Fung, P., Hoi, S.: Instructblip: Towards general-purpose vision-language models with instruction tuning (2023)
7. DeAndres-Tame, I., Tolosana, R., Vera-Rodriguez, R., Morales, A., Fierrez, J., Ortega-Garcia, J.: How good is chatgpt at face biometrics? a first look into recognition, soft biometrics, and explainability (2024)
8. Eidinger, E., Enbar, R., Hassner, T.: Age and gender estimation of unfiltered faces. IEEE Transactions on Information Forensics and Security **9**(12), 2170–2179 (2014). https://doi.org/10.1109/tifs.2014.2359646
9. Fu, T.J., Hu, W., Du, X., Wang, W.Y., Yang, Y., Gan, Z.: Guiding instruction-based image editing via multimodal large language models. In: International Conference on Learning Representations (ICLR) (2024)
10. Gu, X., Lin, T.Y., Kuo, W., Cui, Y.: Open-vocabulary object detection via vision and language knowledge distillation (2021)
11. Hiba, S., Keller, Y.: Hierarchical attention-based age estimation and bias estimation (2021)
12. 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. pp. 1548–1558 (2021)
13. Kipf, T.N., Welling, M.: Semi-supervised classification with graph convolutional networks (2016)
14. Kuprashevich, M., Tolstykh, I.: Mivolo: Multi-input transformer for age and gender estimation (2023)
15. Kuznetsova, A., Rom, H., Alldrin, N., Uijlings, J., Krasin, I., Pont-Tuset, J., Kamali, S., Popov, S., Malloci, M., Kolesnikov, A., Duerig, T., Ferrari, V.: The open images dataset v4: Unified image classification, object detection, and visual relationship detection at scale (2018). https://doi.org/10.1007/s11263-020-01316-z

16. Lai, X., Tian, Z., Chen, Y., Li, Y., Yuan, Y., Liu, S., Jia, J.: Lisa: Reasoning segmentation via large language model (2023)
17. Levi, G., Hassncer, T.: Age and gender classification using convolutional neural networks. In: 2015 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW). pp. 34–42 (2015). https://doi.org/10.1109/CVPRW.2015.7301352
18. Li, C., Wong, C., Zhang, S., Usuyama, N., Liu, H., Yang, J., Naumann, T., Poon, H., Gao, J.: Llava-med: Training a large language-and-vision assistant for biomedicine in one day. arXiv preprint arXiv:2306.00890 (2023)
19. Li, J., Li, D., Savarese, S., Hoi, S.: Blip-2: Bootstrapping language-image pre-training with frozen image encoders and large language models (2023)
20. Li, W., Huang, X., Zhu, Z., Tang, Y., Li, X., Zhou, J., Lu, J.: Ordinalclip: Learning rank prompts for language-guided ordinal regression. In: Koyejo, S., Mohamed, S., Agarwal, A., Belgrave, D., Cho, K., Oh, A. (eds.) Advances in Neural Information Processing Systems. vol. 35, pp. 35313–35325. Curran Associates, Inc. (2022), https://proceedings.neurips.cc/paper_files/paper/2022/file/e55b33430e344a1ee23710415b1c9d87-Paper-Conference.pdf
21. Lin, Y., Shen, J., Wang, Y., Pantic, M.: Fp-age: Leveraging face parsing attention for facial age estimation in the wild. arXiv (2021)
22. Lin, Y., Shen, J., Wang, Y., Pantic, M.: Fp-age: Leveraging face parsing attention for facial age estimation in the wild. arXiv (2021)
23. Liu, H., Li, C., Li, Y., Lee, Y.J.: Improved baselines with visual instruction tuning (2023)
24. Liu, H., Li, C., Li, Y., Li, B., Zhang, Y., Shen, S., Lee, Y.J.: Llava-next: Improved reasoning, ocr, and world knowledge (January 2024), https://llava-vl.github.io/blog/2024-01-30-llava-next/
25. Liu, H., Li, C., Wu, Q., Lee, Y.J.: Visual instruction tuning. In: NeurIPS (2023)
26. Liu, X., Li, S., Kan, M., Zhang, J., Wu, S., Liu, W., Han, H., Shan, S., Chen, X.: Agenet: Deeply learned regressor and classifier for robust apparent age estimation. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) Workshops (December 2015)
27. Ma, H., Zhao, H., Lin, Z., Kale, A., Wang, Z., Yu, T., Gu, J., Choudhary, S., Xie, X.: Ei-clip: Entity-aware interventional contrastive learning for e-commerce cross-modal retrieval. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 18030–18040 (2022). https://doi.org/10.1109/CVPR52688.2022.01752
28. Maaz, M., Rasheed, H., Khan, S., Khan, F.S.: Video-chatgpt: Towards detailed video understanding via large vision and language models. arXiv:2306.05424 (2023)
29. OpenAI: ChatGPT: A Large Language Model. Online; accessed February 13, 2024 (2023), available at https://www.openai.com/
30. OpenAI: Gpt-4 technical report (2023)
31. Paplham, J., Franc, V.: A call to reflect on evaluation practices for age estimation: Comparative analysis of the state-of-the-art and a unified benchmark (2023)
32. Peng, Z., Wang, W., Dong, L., Hao, Y., Huang, S., Ma, S., Wei, F.: Kosmos-2: Grounding multimodal large language models to the world (2023)
33. Radford, A., Kim, J.W., Hallacy, C., Ramesh, A., Goh, G., Agarwal, S., Sastry, G., Askell, A., Mishkin, P., Clark, J., Krueger, G., Sutskever, I.: Learning transferable visual models from natural language supervision. In: Meila, M., Zhang, T. (eds.) Proceedings of the 38th International Conference on Machine Learning. Proceedings of Machine Learning Research, vol. 139, pp. 8748–8763. PMLR (18–24 Jul 2021), https://proceedings.mlr.press/v139/radford21a.html
34. Rao, Y., Zhao, W., Chen, G., Tang, Y., Zhu, Z., Huang, G., Zhou, J., Lu, J.: Denseclip: Language-guided dense prediction with context-aware prompting. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 18061–18070 (2022). https://doi.org/10.1109/CVPR52688.2022.01755
35. Savchenko, A.V.: Efficient facial representations for age, gender and identity recognition in organizing photo albums using multi-output cnn (2018). https://doi.org/10.7717/peerj-cs.197
36. Savchenko, A.V.: Facial expression and attributes recognition based on multi-task learning of lightweight neural networks. In: 2021 IEEE 19th International Symposium on Intelligent Systems and Informatics (SISY). pp. 119–124 (2021). https://doi.org/10.1109/SISY52375.2021.9582508
37. Schuhmann, C., Beaumont, R., Vencu, R., Gordon, C., Wightman, R., Cherti, M., Coombes, T., Katta, A., Mullis, C., Wortsman, M., Schramowski, P., Kundurthy, S., Crowson, K., Schmidt, L., Kaczmarczyk, R., Jitsev, J.: Laion-5b: An open large-scale dataset for training next generation image-text models (2022)
38. Shin, N.H., Lee, S.H., Kim, C.S.: Moving window regression: A novel approach to ordinal regression. In: 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 18739–18748 (2022). https://doi.org/10.1109/CVPR52688.2022.01820
39. Shou, Y.W., Cao, X., Meng, D.: Masked contrastive graph representation learning for age estimation. ArXiv **abs/2306.17798** (2023), https://api.semanticscholar.org/CorpusID:259309260

40. Shou, Y., Ai, W., Meng, T., Li, K.: Czl-ciae: Clip-driven zero-shot learning for correcting inverse age estimation (2023)
41. Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., Rodriguez, A., Joulin, A., Grave, E., Lample, G.: Llama: Open and efficient foundation language models (2023)
42. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, L.u., Polosukhin, I.: Attention is all you need. In: Guyon, I., Luxburg, U.V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 30. Curran Associates, Inc. (2017), https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf
43. Wang, H., Sanchez, V., Li, C.T.: Improving face-based age estimation with attention-based dynamic patch fusion. IEEE Transactions on Image Processing **31**, 1084–1096 (2022). https://doi.org/10.1109/TIP.2021.3139226
44. Wang, R., Li, P., Huang, H., Cao, C., He, R., He, Z.: Learning-to-rank meets language: Boosting language-driven ordering alignment for ordinal classification (2023)
45. Wang, W., Chen, Z., Chen, X., Wu, J., Zhu, X., Zeng, G., Luo, P., Lu, T., Zhou, J., Qiao, Y., Dai, J.: Visionllm: Large language model is also an open-ended decoder for vision-centric tasks (2023)
46. Ye, Q., Xu, H., Xu, G., Ye, J., Yan, M., Zhou, Y., Wang, J., Hu, A., Shi, P., Shi, Y., Li, C., Xu, Y., Chen, H., Tian, J., Qi, Q., Zhang, J., Huang, F.: mplug-owl: Modularization empowers large language models with multimodality (2023)
47. Zhang, C., Liu, S., Xu, X., Zhu, C.: C3ae: Exploring the limits of compact model for age estimation. In: 2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR). pp. 12579–12588 (2019). https://doi.org/10.1109/CVPR.2019.01287
48. Zhang, Y., Liu, L., Li, C., change Loy, C.: Quantifying facial age by posterior of age comparisons (2017)
49. Zhu, D., Chen, J., Shen, X., Li, X., Elhoseiny, M.: Minigpt-4: Enhancing vision-language understanding with advanced large language models (2023)

Authors

Maksim Kuprashevich received a Bachelor's degree in Computer Science from the Saint-Petersburg State Institute of Technology. He is currently a Research Team Lead. His research interests include deep learning, particularly in vision, language, and generative models.

Grigorii Alekseenko received a Specialist degree in Fundamental Mathematics and Mechanics from Moscow State University. He is currently a Data Scientist. His research interests include computer vision, multimodality, and diffusion neural networks.

Irina Tolstykh received a Bachelor's degree in Fundamental Informatics and Information Technology from Saint-Petersburg State University. She is currently a Senior Data Scientist. Her research interests in machine learning include applying deep learning methods to computer vision and natural language processing, as well as exploring generative AI.