# BENCHMARKING MULTIMODAL LARGE LANGUAGE MODELS FOR FACE RECOGNITION

*Hatef Otroshi Shahreza and Sébastien Marcel*

Idiap Research Institute, Switzerland
{hatef.otroshi, sebastien.marcel}@idiap.ch

## ABSTRACT

Multimodal large language models (MLLMs) have achieved remarkable performance across diverse vision-and-language tasks. However, their potential in face recognition remains underexplored. In particular, the performance of open-source MLLMs needs to be evaluated and compared with existing face recognition models on standard benchmarks with similar protocol. In this work, we present a systematic benchmark of state-of-the-art MLLMs for face recognition on several face recognition datasets, including LFW, CALFW, CPLFW, CFP, AgeDB and RFW. Experimental results reveal that while MLLMs capture rich semantic cues useful for face-related tasks, they lag behind specialized models in high-precision recognition scenarios in zero-shot applications. This benchmark provides a foundation for advancing MLLM-based face recognition, offering insights for the design of next-generation models with higher accuracy and generalization. The source code of our benchmark is publicly available.

***Index Terms*—** Benchmark, Face Recognition, Foundation Models, Multimodal Large Language Models (MLLMs)

## 1. INTRODUCTION

Multimodal large language models (MLLMs) have recently gained significant attention from the research community for visual and linguistic understanding tasks. By combining pretrained visual encoders with large language models (LLMs), systems such as Flamingo [1], QwenVL [2], and GPT-4o [3] have achieved state-of-the-art performance across diverse tasks, including image captioning and visual question answering (VQA). These models showcase the ability of LLMs to reason over perceptual inputs and generate coherent, contextually grounded output text, enabling general-purpose image processing in zero-shot and few-shot settings. Leveraging large-scale pretraining, they have accelerated the development of foundation models that are capable of interpreting and responding to complex visual questions without requiring task-specific supervision.

This work was funded by the Hasler foundation through the Responsible Face Recognition (SAFER) project and also by the European Union project CarMen (Grant Agreement No. 101168325).

Face recognition is also a popular computer vision task and is increasingly used in different applications [4, 5, 6]. In particular, face recognition is used as a secure authentication tool in a broad range of applications such as smart phone unlocking, border control, etc. In addition to the security purposes, face recognition is used for entertainment and also in social media. Face recognition models have been extensively studied in the literature and there are also standard benchmarks to evaluate and compare the performance of face recognition models.

With the surge of MLLMs, we can consider potential applications of MLLMs for face recognition [7]. However, for replacing MLLMs with existing face recognition models, it is important to know the performance of MLLMs compared to typical models on standard benchmarks with similar protocol. In this paper, we investigate *how open-source MLLMs perform on face recognition benchmarks?* While there are some previous works on evaluation of MLLMs for different face understanding tasks [8, 9, 10], to our knowledge this paper is the first work that benchmarks MLLMs for face recognition on standard datasets with similar protocols.

In the remaining of this paper, we first review previous work in the literature in Section 2. We then describe our benchmark in Section 3 and present our results in Section 4. Finally, the paper is concluded in Section 5.

## 2. RELATED WORK

Recently, several papers have investigated the application of MLLMs for face-related tasks, including face recognition, attribute analysis, forgery detection, anti-spoofing, and multimodal reasoning. A recent survey [7] provides an extensive review of how foundation models and MLLMs are being applied in biometrics and face recognition.

Early studies investigated the use of pretrained MLLMs, such as ChatGPT [3], for face verification [9], and predicting soft-biometrics, such as age, gender, and ethnicity. Jia *et al.* [11] also evaluated the application of ChatGPT for zero-shot face deepfake detection. Shi *et al.* [12] explored chain-of-thoughts prompting for Gemini and ChatGPT in face anti-spoofing and deepfake detection tasks. Komaty *et al.* [10]

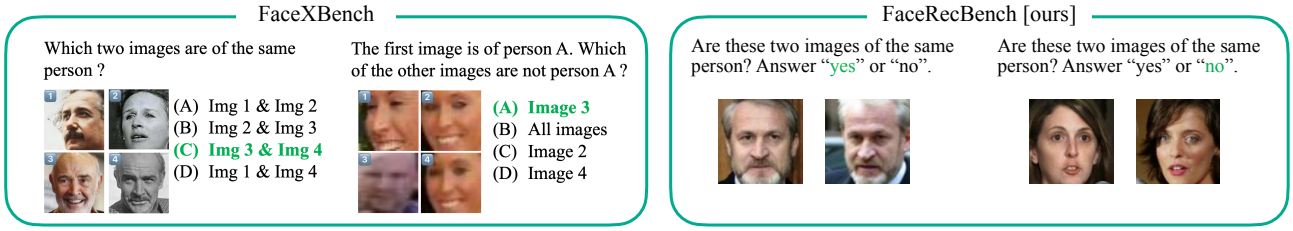


**Fig. 1**. Sample questions in FaceXBench [8] and our benchmark.

investigated in-context learning of ChatGPT [3] for face anti-spoofing. Sony *et al.* [13] evaluated the performance of several foundation models (such as CLIP, BLIP, etc.) for face recognition, and showed that fusion of face recognition models with foundation models can improve recognition accuracy.

In addition to these studies, some benchmarks were proposed for different face processing tasks. FaceBench [14] proposed a visual question-answering benchmark for facial attributes. Benchmarks such as FaceXBench [8] and Face-Human-Bench [15] were also proposed to benchmark MLLMs across various face processing tasks, including facial expression recognition, attribute prediction, anti-spoofing, etc. FaceXBench [8] also includes face recognition, using face recognition datasets such as LFW [16], AgeDB [17], CFP-FF [18], CFP-FP [18], CALFW [19], CPLFW [20]. However, they used multiple-choice questions for evaluating the performance of MLLMs. Fig. 1 illustrates two example questions from FaceXBench for face recognition task. While FaceXBench is a useful benchmark for comparing MLLMs in face processing tasks, the reported accuracy values are not comparable to the accuracy of face recognition models in the literature. In fact, for evaluating typical face recognition models, we simply have two face images and want to see if they have same identity. However, reported values for face recognition based on questions with multiple images and also multiple choices are not consistent with values reported in the literature. In this paper, we focus on face recognition and benchmark MLLMs with similar protocols as typical face recognition models.

## 3. BENCHMARKING MLLMS FOR FACE RECOGNITION

To evaluate MLLMs for face recognition, we consider a verification task where two face images are available and the question is whether the given images belong to the same identity. Hence, we give the MLLM with both images along with the following prompt:

**Prompt**

Are these two images of the same person? Answer "yes" or "no".

Then, the output of MLLM is expected to be "yes" or "no", meaning if the images are predicted to correspond to same person or not.

In our benchmark, we consider different standard datasets, including Labeled Faces in the Wild (LFW) [16], Cross-age LFW (CALFW) [19], Cross-Pose LFW (CPLFW) [20], Celebrities in Frontal-Profile in the Wild (CFP) [18], AgeDB-30 [17], and Racial Faces in-the-Wild (RFW) [21]. Our evaluation for each of these dataset include 6,000 pairs of images with 3,000 positive and 3,000 negative pairs. For consistency with prior works on face recognition, we report recognition accuracy on these datasets. In the following, we briefly describe each dataset:

**Labeled Faces in the Wild (LFW):** LFW [16] is a widely used benchmark dataset for unconstrained face verification. It contains 13,233 images of faces collected from the web, with 5,749 unique individuals. The dataset is designed to evaluate how well face recognition algorithms generalize to real-world conditions, with variations in pose, expression, illumination, and background. Since its release, LFW has served as a standard reference point for measuring progress in face recognition under unconstrained settings.

**Cross-Age LFW (CALFW):** CALFW [19] extends LFW by introducing cross-age variation, aiming to make the verification task more challenging. It contains image pairs of the same individual captured at different ages, highlighting the difficulty of recognizing faces over long time spans. This dataset is primarily focused on age-related intra-class variations while maintaining inter-class diversity, making it a valuable benchmark for studying the robustness of face recognition systems to aging effects.

**Cross-Pose LFW (CPLFW):** CPLFW [20] is another variant of LFW, created to evaluate face recognition under cross-pose conditions. It includes face pairs where the same subject appears in significantly different poses, thus introducing large intra-class variations in pose. By emphasizing pose differences, CPLFW complements LFW and CALFW to how well recognition systems handle extreme viewpoint changes.

**Celebrities in Frontal-Profile (CFP):** The CFP [18] dataset is introduced to test face recognition across frontal and profile views. It consists of images of 500 celebrities, with both frontal and profile face shots, and provides verification protocols for frontal-to-frontal (CFP-FF) and frontal-

**Table 1**. Comparison of recognition accuracy (%) of MLLMs with face recognition models on different face datasets.

| Model | LFW | AgeDB30 | CALFW | CPLFW | CFP-FP | CFP-FF | Average |
|---|---|---|---|---|---|---|---|
| **Open source MLLMs** | | | | | | | |
| LLaVA-v1.5-7b | 50.00 | 50.00 | 50.00 | 50.00 | 50.00 | 50.00 | 50.00 |
| LLaVA-v1.5-13b | 49.92 | 49.68 | 50.02 | 49.83 | 50.09 | 50.01 | 49.93 |
| LLaVA-OneVision-Qwen2-0.5b | 55.52 | 52.63 | 50.08 | 51.55 | 55.00 | 55.61 | 53.40 |
| LLaVA-NeXT-Vicuna-13b | 65.95 | 57.50 | 51.53 | 54.95 | 67.44 | 70.87 | 61.37 |
| LLaVA-NeXT-Vicuna-7b | 53.02 | 50.72 | 50.13 | 50.13 | 53.87 | 56.86 | 52.45 |
| LLaVA-NeXT-Mistral-7b | 50.00 | 49.98 | 50.02 | 50.22 | 49.99 | 50.03 | 50.04 |
| GLM-4v-9b | 52.58 | 50.42 | 48.52 | 50.27 | 50.39 | 49.93 | 50.35 |
| Idefics-9b-Instruct | 50.13 | 50.05 | 50.02 | 49.98 | 49.99 | 50.40 | 50.09 |
| Idefics2-8b | 72.40 | 68.53 | 54.93 | 55.98 | 74.11 | 74.43 | 66.73 |
| Idefics3-8B-Llama3 | 88.83 | 57.98 | 61.18 | 70.90 | 80.26 | 85.11 | 74.05 |
| ShareGPT4v-7b | 49.98 | 50.00 | 50.00 | 49.95 | 50.00 | 50.00 | 49.99 |
| ShareGPT4v-13b | 49.92 | 49.98 | 50.00 | 49.87 | 50.16 | 50.00 | 49.99 |
| PaliGemma-3b-mix-448 | 48.48 | 49.40 | 48.43 | 50.37 | 50.26 | 50.33 | 49.54 |
| Ovis1.5-Llama3-8B | 52.63 | 52.40 | 50.85 | 50.23 | 53.67 | 54.97 | 52.46 |
| Ovis1.5-Gemma2-9B | 73.93 | 57.87 | 56.95 | 54.93 | 75.09 | 78.49 | 66.21 |
| Llama-3.2-11B-Vision-Instruct | 50.55 | 48.20 | 51.83 | 49.50 | 49.47 | 49.94 | 49.92 |
| InternVL2.5-1B | 54.22 | 50.62 | 50.80 | 51.10 | 51.44 | 51.26 | 51.57 |
| InternVL3-1B | 69.28 | 56.25 | 56.15 | 63.35 | 60.80 | 64.06 | 61.65 |
| InternVL3-8B | 87.92 | 52.27 | 59.08 | 72.30 | 79.20 | 81.84 | 72.10 |
| InternVL3-38B | 90.10 | 55.72 | 61.37 | 71.20 | 72.50 | 84.40 | 72.55 |
| FaceLLM-8B | 90.65 | 53.38 | 61.48 | 73.50 | 80.06 | 84.89 | 73.99 |
| Valley2 | 92.93 | 60.75 | 68.58 | 74.55 | 84.33 | 92.27 | 78.90 |
| Qwen2-VL-2B-Instruct | 63.38 | 58.55 | 50.77 | 51.17 | 61.27 | 62.33 | 57.91 |
| Qwen2-VL-7B-Instruct | **93.28** | **66.03** | **71.95** | **75.28** | **86.93** | **93.11** | **81.10** |
| Qwen2.5-VL-3B-Instruct | 77.52 | 54.03 | 60.92 | 59.43 | 67.00 | 82.70 | 66.93 |
| Qwen2.5-VL-7B-Instruct | 89.48 | 59.07 | 69.08 | 73.43 | 79.09 | 89.46 | 76.60 |
| Qwen2.5-VL-32B-Instruct | 79.32 | 59.07 | 64.48 | 66.15 | 69.70 | 83.01 | 70.29 |
| **Face Recognition Models** | | | | | | | |
| IResNet-50 (HyperFace) | 98.27 | 90.40 | 91.48 | 85.60 | 92.24 | 98.86 | 92.81 |
| IResNet-50 (MS1MV2) | **99.83** | **98.28** | **95.45** | **92.08** | **98.27** | **99.99** | **97.31** |

to-profile (CFP-FP) matching.

**AgeDB:** AgeDB [17] is a benchmark dataset focused on age-related variations in face recognition. It contains images 16,516 images of 570 subjects with a wide age range. The dataset provides predefined verification protocols with increasing age gaps (e.g., 5, 10, 20, and 30 years), enabling systematic evaluation of how well algorithms handle the challenge of age progression. AgeDB is commonly used to study long-term face recognition performance. We use 30-year protocol in our benchmark.

**Racial Faces in-the-Wild (RFW):** The RFW [21] dataset was introduced to evaluate bias and fairness in face recognition systems across different demographic groups. RFW is constructed by reorganizing images from MS-Celeb [22] into four balanced subsets: Caucasian, Asian, Indian, and African. Each subset contains approximately 10,000 images from around 3,000 individuals, with 6,000 comparisons. By providing a benchmark focused on racial diversity, RFW enables systematic analysis of demographic disparities in recognition performance and has become a widely used dataset for studying fairness in face recognition.

We use the cropped images for each dataset available in Insightface repository[1]. We also use VLMEvalKit repository[2] to implement our benchmark. The source code of our benchmark is publicly available in the project page[3].

## 4. EXPERIMENTAL RESULTS

We evaluate and benchmark open-source[4] MLLMs on various standard face recognition datasets, including LFW [16], CALFW [19], CPLFW [20], CFP [18], AgeDB-30 [17], and RFW [21]. The MLLMs used in our experiments include LLaVA-v1.5-7b [23], LLaVA-v1.5-13b [23],

[1]https://github.com/deepinsight/insightface

[2]https://github.com/open-compass/VLMEvalKit

[3]Project page: https://www.idiap.ch/paper/facerecbench

[4]Note that given the restrictions in license of each of the benchmark datasets, we were not able to use commercial MLLMs in this study.

**Table 2**. Comparison of recognition accuracy (%) of MLLMs on different demography groups in RFW.

| Model | African | Asian | Caucasian | Indian | Average | Std. |
|---|---|---|---|---|---|---|
| **Open source MLLMs** | | | | | | |
| LLaVA-NeXT-Vicuna-13b | 51.28 | 53.53 | 56.88 | 53.00 | 53.67 | 2.03 |
| Idefics3-8B-Llama3 | 60.78 | 66.15 | 70.38 | 66.38 | 65.92 | 3.41 |
| Ovis1.5-Gemma2-9B | 52.62 | 54.92 | 61.93 | 55.83 | 56.33 | 3.44 |
| InternVL3-8B | 64.37 | 63.08 | 66.37 | 64.03 | 64.46 | 1.20 |
| FaceLLM-8B | **66.00** | 66.78 | 68.82 | 66.02 | 66.90 | **1.15** |
| Valley2 | 63.53 | **68.77** | 75.57 | **70.28** | **69.54** | 4.29 |
| Qwen2-VL-7B-Instruct | 60.40 | 65.55 | **76.68** | 68.35 | 67.75 | 5.90 |
| Qwen2.5-VL-7B-Instruct | 62.65 | 66.63 | 70.55 | 68.98 | 67.20 | 2.98 |
| **Face Recognition Models** | | | | | | |
| IResNet-50 (HyperFace) | 88.27 | 82.98 | 84.33 | 78.02 | 83.40 | 3.66 |
| IResNet-50 (MS1MV2) | **98.32** | **97.73** | **99.33** | **98.23** | **98.40** | **0.58** |

LLaVA-OneVision-Qwen2-0.5b [24], LLaVA-NeXT-Vicuna-7b [24], LLaVA-NeXT-Vicuna-13b [24], LLaVA-NeXT-Mistral-7b [24], GLM-4v-9b [25], Idefics-2-8b [26], Idefics-9b-Instruct [27], Idefics3-8B-Llama3 [28], ShareGPT4v-13b [29], ShareGPT4v-7b [29], PaliGemma-3b-mix-448 [30], Ovis-1.5-Llama3-8B [31], Ovis1.5-Gemma2-9B [31], Llama-3.2-11B-Vision-Instruct [32], InternVL2.5-1B [33], InternVL3-1B [34], InternVL3-2B [34], InternVL3-8B [34], FaceLLM-8B [35], Valley2 [36], Qwen2-VL-2B-Instruct [2], Qwen2-VL-7B-Instruct [2], Qwen2.5-VL-3B-Instruct [37], Qwen2.5-VL-7B-Instruct [37], Qwen2.5-VL-32B-Instruct [37]. We run our evaluations on a system equipped with NVIDIA H100.

Table 1 reports the performance of different MLLMs on several face recognition benchmarks (LFW, CALFW, CPLFW, CFP, and AgeDB-30). This table also compares the performance of MLLMs with IResNet-50 (trained with AdaFace loss [6] on MS-Celeb [22] dataset) as a state-of-the-art face recognition model. As another baseline, we also consider a face recognition model with IResNet-50 trained with HyperFace [38] synthetic dataset. As the results in this paper show there is a significant gap between the performance of face recognition models. While increasing the size of MLLM can improve the performance on benchmarks, it also saturates for each MLLM (can be seen in InternVL3 and Qwen2.5VL).

Among the benchmarked MLLMs, FaceLLM is based on InternVL3 and finetuned for face understanding. The results in Table 1 show that the finetuing in FaceLLM has increased the performance compared to the base model (InternVL3) on different face recognition benchmarks. This suggests that by using domain-specific data instead of general-purpose data, we can expect improvement in MLLMs for the face recognition task.

We also compare the top-performing models in Table 1 on RFW dataset. Table 2 reports the performance of different models for four demography groups. The result in this table also indicate significant gap between the performance of MLLMs with typical models.

## 5. CONCLUSION

In this paper, we presented a benchmark for MLLMs with similar protocol used to evaluate typical face recognition models. Although MLLMs have shown considerable potentials in broad applications, most are trained mainly on general-purpose datasets or large-scale image–text pairs collected from the web. Consequently, these models are able to generate high-level image descriptions but often lack task-specific precision. For instance, while they can describe a person's appearance or identify basic demographic attributes such as age and gender, they frequently struggle with more details which are required to recognize identity or verify if identity is the same in two images. This limitation poses challenges for applications of MLLMs in face recognition and requires further study in future. Our benchmark can be used by future researchers to compare the MLLMs with face recognition models.

## 6. REFERENCES


[1] Jean-Baptiste Alayrac et al., "Flamingo: a visual language model for few-shot learning," *Advances in neural information processing systems*, vol. 35, pp. 23716–23736, 2022.

[2] Peng Wang et al., "Qwen2-vl: Enhancing vision-language model's perception of the world at any resolution," *arXiv preprint arXiv:2409.12191*, 2024.

[3] Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, et al., "Gpt-4o system card," *arXiv preprint arXiv:2410.21276*, 2024.

[4] Jiankang Deng, Jia Guo, Niannan Xue, and Stefanos Zafeiriou, "Arcface: Additive angular margin loss for deep face recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2019, pp. 4690–4699.

[5] Anjith George, Christophe Ecabert, Hatef Otroshi Shahreza, Ketan Kotwal, and Sébastien Marcel, "Edgeface: Efficient face recognition model for edge devices," *IEEE Transactions on Biometrics, Behavior, and Identity Science*, vol. 6, no. 2, pp. 158–168, 2024.

[6] Minchul Kim, Anil K Jain, and Xiaoming Liu, "Adaface: Quality adaptive margin for face recognition," in *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, 2022, pp. 18750–18759.

[7] Hatef Otroshi Shahreza and Sébastien Marcel, "Foundation models and biometrics: A survey and outlook," *IEEE Transactions on Information Forensics and Security*, 2025.

[8] Kartik Narayan, Vibashan VS, and Vishal M Patel, "Facexbench: Evaluating multimodal llms on face understanding," *arXiv preprint arXiv:2501.10360*, 2025.

[9] Ahmad Hassanpour, Yasamin Kowsari, Hatef Otroshi Shahreza, Bian Yang, and Sébastien Marcel, "Chatgpt and biometrics: an assessment of face recognition, gender detection, and age estimation capabilities," in *2024 IEEE International Conference on Image Processing (ICIP)*. IEEE, 2024, pp. 3224–3229.

[10] Alain Komaty, Hatef Otroshi Shahreza, Anjith George, and Sebastien Marcel, "Exploring chatgpt for face presentation attack detection in zero and few-shot in-context learning," *arXiv preprint arXiv:2501.08799*, 2025.

[11] Shan Jia, Reilin Lyu, Kangran Zhao, Yize Chen, Zhiyuan Yan, Yan Ju, Chuanbo Hu, Xin Li, Baoyuan Wu, and Siwei Lyu, "Can chatgpt detect deepfakes? a study of using multimodal large language models for media forensics," in *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2024, pp. 4324–4333.

[12] Yichen Shi et al., "Shield: An evaluation benchmark for face spoofing and forgery detection with multimodal large language models," *arXiv preprint arXiv:2402.04178*, 2024.

[13] Redwan Sony, Parisa Farmanifard, Arun Ross, and Anil K Jain, "Foundation versus domain-specific models: Performance comparison, fusion, and explainability in face recognition," *arXiv preprint arXiv:2507.03541*, 2025.

[14] Xiaoqin Wang, Xusen Ma, Xianxu Hou, Meidan Ding, Yudong Li, Junliang Chen, Wenting Chen, Xiaoyang Peng, and Linlin Shen, "Facebench: A multi-view multi-level facial attribute vqa dataset for benchmarking face perception mllms," *arXiv preprint arXiv:2503.21457*, 2025.

[15] Lixiong Qin et al., "Face-human-bench: A comprehensive benchmark of face and human understanding for multi-modal assistants," *arXiv preprint arXiv:2501.01243*, 2025.

[16] Gary B Huang, Marwan Mattar, Tamara Berg, and Eric Learned-Miller, "Labeled faces in the wild: A database forstudying face recognition in unconstrained environments," in *Workshop on faces in'Real-Life'Images: detection, alignment, and recognition*, 2008.

[17] Stylianos Moschoglou, Athanasios Papaioannou, Christos Sagonas, Jiankang Deng, Irene Kotsia, and Stefanos Zafeiriou, "Agedb: the first manually collected, in-the-wild age database," in *proceedings of the IEEE conference on computer vision and pattern recognition workshops*, 2017, pp. 51–59.

[18] Soumyadip Sengupta, Jun-Cheng Chen, Carlos Castillo, Vishal M Patel, Rama Chellappa, and David W Jacobs, "Frontal to profile face verification in the wild," in *2016 IEEE winter conference on applications of computer vision (WACV)*. IEEE, 2016, pp. 1–9.

[19] Tianyue Zheng, Weihong Deng, and Jiani Hu, "Cross-age lfw: A database for studying cross-age face recognition in unconstrained environments," *arXiv preprint arXiv:1708.08197*, 2017.

[20] Tianyue Zheng and Weihong Deng, "Cross-pose lfw: A database for studying cross-pose face recognition in unconstrained environments," *Beijing University of Posts and Telecommunications, Tech. Rep*, vol. 5, no. 7, 2018.

[21] Mei Wang, Weihong Deng, Jiani Hu, Xunqiang Tao, and Yaohai Huang, "Racial faces in the wild: Reducing racial bias by information maximization adaptation network," in *Proceedings of the ieee/cvf international conference on computer vision*, 2019, pp. 692–702.

[22] Yandong Guo, Lei Zhang, Yuxiao Hu, Xiaodong He, and Jianfeng Gao, "Ms-celeb-1m: A dataset and benchmark for large-scale face recognition," in *European conference on computer vision*. Springer, 2016, pp. 87–102.

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

[24] Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee, "Improved baselines with visual instruction tuning," 2023.

[25] Team GLM, Aohan Zeng, et al., "Chatglm: A family of large language models from glm-130b to glm-4 all tools," 2024.

[26] Hugo Laurençon, Léo Tronchon, Matthieu Cord, and Victor Sanh, "What matters when building vision-language models?," 2024.

[27] Hugo Laurençon, Lucile Saulnier, Léo Tronchon, Stas Bekman, Amanpreet Singh, Anton Lozhkov, Thomas Wang, Siddharth Karamcheti, Alexander M. Rush, Douwe Kiela, Matthieu Cord, and Victor Sanh, "Obelics: An open web-scale filtered dataset of interleaved image-text documents," 2023.

[28] Hugo Laurençon, Andrés Marafioti, Victor Sanh, and Léo Tronchon, "Building and better understanding vision-language models: insights and future directions.," 2024.

[29] Lin Chen and other, "Sharegpt4v: Improving large multimodal models with better captions," in *European Conference on Computer Vision*. Springer, 2024, pp. 370–387.

[30] Lucas Beyer, Andreas Steiner, André Susano Pinto, Alexander Kolesnikov, Xiao Wang, Daniel Salz, Maxim Neumann, Ibrahim Alabdulmohsin, Michael Tschannen, Emanuele Bugliarello, et al., "Paligemma: A versatile 3b vlm for transfer," *arXiv preprint arXiv:2407.07726*, 2024.

[31] Shiyin Lu, Yang Li, Qing-Guo Chen, Zhao Xu, Weihua Luo, Kaifu Zhang, and Han-Jia Ye, "Ovis: Structural embedding alignment for multimodal large language model," *arXiv:2405.20797*, 2024.

[32] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al., "The llama 3 herd of models," *arXiv preprint arXiv:2407.21783*, 2024.

[33] Zhe Chen et al., "Expanding performance boundaries of open-source multimodal models with model, data, and test-time scaling," *arXiv preprint arXiv:2412.05271*, 2024.

[34] Jinguo Zhu et al., "Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models," *arXiv preprint arXiv:2504.10479*, 2025.

[35] Hatef Otroshi Shahreza and Sébastien Marcel, "Facellm: A multimodal large language model for face understanding," *arXiv preprint arXiv:2507.10300*, 2025.

[36] Ziheng Wu, Zhenghao Chen, Ruipu Luo, Can Zhang, Yuan Gao, Zhentao He, Xian Wang, Haoran Lin, and Minghui Qiu, "Valley2: Exploring multimodal models with scalable vision-language design," *arXiv preprint arXiv:2501.05901*, 2025.

[37] Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, et al., "Qwen2.5 technical report," 2025.

[38] Hatef Otroshi Shahreza and Sébastien Marcel, "Hyperface: Generating synthetic face recognition datasets by exploring face embedding hypersphere," in *The Thirteenth International Conference on Learning Representations*, 2025.