Evaluating Multimodal Large Language Models for Heterogeneous Face Recognition

Hatef Otroshi Shahreza, Anjith George, Sébastien Marcel
Idiap Research Institute, Switzerland
{hatef.otroshi,anjith.george,sebastien.marcel}@idiap.ch

Abstract

Multimodal Large Language Models (MLLMs) have recently demonstrated strong performance on a wide range of vision-language tasks, raising interest in their potential use for biometric applications. In this paper, we conduct a systematic evaluation of state-of-the-art MLLMs for heterogeneous face recognition *(HFR), where enrollment and probe images are from different sensing modalities, including visual (VIS), near infrared (NIR), short-wave infrared (SWIR), and thermal camera. We benchmark multiple open-source MLLMs across several cross-modality scenarios, including VIS-NIR, VIS-SWIR, and VIS-THERMAL face recognition. The recognition performance of MLLMs is evaluated using biometric protocols and based on different metrics, including Acquire Rate, Equal Error Rate (EER), and True Accept Rate (TAR). Our results reveal substantial performance gaps between MLLMs and classical face recognition systems, particularly under challenging cross-spectral conditions, in spite of recent advances in MLLMs. Our findings highlight the limitations of current MLLMs for HFR and also the importance of rigorous biometric evaluation when considering their deployment in face recognition systems.*

1. Introduction

Face recognition has achieved remarkable progress over the past decade, driven largely by deep convolutional and transformer-based models trained on large-scale visual datasets [8, 11]. Existing face recognition systems often perform reliably in *homogeneous* settings, where enrollment and probe images are captured under the same modality. However, the performance of such models degrade significantly in *heterogeneous face recognition* (HFR) scenarios, where enrollment and probe images are from different sensing modalities. Heterogeneous face recognition is particularly important, where we do not have enough data to train a model for recognition, such as new sensing cameras or special environmental conditions. Examples of such applications include surveillance, border control, low-light environments, etc.

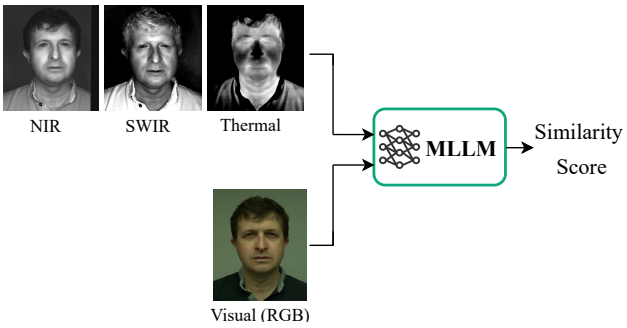

Figure 1. Heterogeneous face recognition using Multimodal Large Language Models (MLLMs)

In recent years, Multimodal Large Language Models (MLLMs) have emerged as a popular and powerful subset of foundation models that are able to process visual and textual inputs. Recent MLLMs demonstrate significant performance in zero-shot and few-shot scenarios on different vision-language tasks, including image captioning, visual question answering, and multimodal reasoning. These capabilities have led to growing interest in whether MLLMs can be directly applied to different vision tasks, such as face recognition, and potentially offering a unified framework that provides perception and reasoning [40, 52, 2].

However, face recognition, particularly in heterogeneous settings, differs fundamentally from many vision-language tasks. Biometric recognition requires fine-grained identity discrimination, robustness to sensing variations, and adherence to strict evaluation protocols. HFR is specially a challenging task, since data from two different sensor types are compared and require strong perception and reasoning to match both data types. While several recent studies have explored the use of large vision or multimodal models for face-related tasks [40, 43, 15, 34, 39], their effectiveness for HFR remains unexplored. In particular, it is unclear how well MLLMs can generalize across sensing modalities and how their performance compares to established face recognition baselines.

In this paper, we present an empirical evaluation of state-of-the-art MLLMs for heterogeneous face recognition. We focus on assessing existing open-source MLLMs under standardized biometric evaluation protocols. We consider different sensing modalities, including visual (VIS), near infrared (NIR), short-wave infrared (SWIR), and thermal camera, and evaluate multiple MLLMs across diverse cross-spectral scenarios, including VIS-NIR, VIS-SWIR, and VIS-THERMAL (Fig. 1). We also compare the performance of MLLMs with face recognition models as reference baselines. To our knowledge, this paper presents the first evaluation of MLLMs for heterogeneous face recognition.

While MLLMs are pretrained on large amount of data, the vast majority of available image data are visual (RGB) images. Our experimental results indicate that, despite their strong multimodal reasoning capabilities, current MLLMs struggle to match the performance of dedicated face recognition models in heterogeneous settings. Performance degradation is particularly pronounced in challenging cross-spectral scenarios, underscoring the gap between general-purpose multimodal understanding and biometric identity recognition. Our findings provide insights into the current limitations of MLLMs for domain change, such as in HFR, and outline important considerations for future research at the intersection of foundation models and biometric recognition.

The remainder of the paper is organized as follows. In Section 2, we review related work in the literature. In Section 3, we explain our benchmark and present our results in Section 4. Finally, the paper is concluded in Section 5.

2. Related Work

2.1. Heterogeneous Face Recognition

The main challenge in HFR is the "domain gap", the significant differences in appearance between images taken with different sensors. This difference makes it hard to accurately match faces and reduces recognition performance. Several approaches have been proposed in the literature to address this challenge.

Common-space projection methods aim to reduce the domain gap by learning mappings that project heterogeneous face representations into a shared subspace [21, 16]. Early studies relied on linear techniques such as CCA [50], PLS [41], and regression-based mappings [25, 26] to align facial features across modalities like NIR and VIS, while discriminative and prototype-based subspace learning further improved cross-modal recognition [30, 23]. More recent deep learning approaches utilize this concept into CNN architectures, for instance through domain-specific units that adapt low-level features across sensing modalities [7], or discriminator-based frameworks that explicitly learn cross-modality relationships using metric learning objectives [4].

Invariant feature–based methods focus on extracting modality-agnostic representations that remain stable across heterogeneous sensing conditions enabling cross-modal face matching. Early approaches emphasized handcrafted descriptors, such as DoG type representations combined with multi-scale LBP features to capture structural consistency across domains [29, 28]. Local feature–driven strategies further incorporated SIFT and MLBP descriptors within discriminative subspace learning frameworks to address modality gap [22, 33, 35]. Information-theoretic methods such as coupled encoding schemes designed to maximize mutual information between heterogeneous modalities [53] were also used for HFR. More recently, CNN-based approaches have emerged to learn invariant facial representations directly from data, demonstrating improved generalization across modalities in HFR tasks [16, 17]. Recently, xEdgeFace [13] introduced a lightweight HFR framework that adapts a pre-trained CNN-Transformer face recognition backbone [11] to cross-modal settings by selectively updating Layer Normalization and early layers within a contrastive self-distillation setting. This strategy effectively aligns heterogeneous modalities while preserving RGB recognition performance, achieving state-of-the-art results across multiple HFR benchmarks with very low computational overhead.

Synthesis-based approaches for HFR aim to reduce modality gap by transforming images from the target domain into the source domain (typically visible spectrum images), allowing recognition with conventional face recognition models [45, 10]. Early methods relied on patch-based or manifold learning techniques for cross-modal image generation [49, 32], while recent advances leverage GAN-based frameworks to synthesize identity-preserving visible faces from heterogeneous modalities, significantly improving recognition performance [55, 51, 10, 12]. However, these methods are typically computationally heavy due to the inclusion of an additional image generation stage.

2.2. MLLMs and Biometrics

Recently, several papers have explored applications of MLLMs for biometrics, including biometric recognition, attribute analysis, forgery detection, presentation attack detection, etc. A recent survey [40] review how MLLMs are being applied in biometrics recognition and security.

Early studies investigated the use of pretrained MLLMs, such as ChatGPT [19], for biometric recognition (e.g., face [15] or iris recognition [9]) and attribute prediction. Jia *et al.* [20] evaluated the performance of ChatGPT for zero-shot face deepfake detection. Komaty *et al.* [24] explored in-context learning using ChatGPT [19] for face presentation attack detection. Shi *et al.* [42] used chain-of-thoughts

prompting with ChatGPT and Gemini for face deepfake detection and presentation attack detection tasks. Sony *et al.* [44] explored foundation models (such as CLIP, BLIP, etc.) for face recognition.

Several benchmarks were also proposed for evaluating MLLMs for face processing tasks. Gavas *et al.* proposed FPBench for benchmarking MLLMs for fingerprint recognition. Wang *et al.* proposed FaceBench [48] as a visual question-answering benchmark for facial attribute prediction with MLLMs. Similarly, other benchmarks, such as FaceXBench [34], FaceRecBench [38] and Face-Human-Bench [37] were proposed to benchmark MLLMs across various face processing tasks, including face recognition, facial expression recognition, attribute prediction, anti-spoofing, etc. FaceXBench [34] and FaceRecBench [38] include face recognition benchmark based on face recognition datasets. FaceXBench [34] used multiple-choice questions for evaluating the performance of MLLMs, while FaceRecBench [38] used Yes/No questions for matching pairs of images. Ozturk *et al.* [36] also evaluated demographic fairness of MLLMs for face recognition. In this paper, we focus on heterogeneous face recognition and benchmark MLLMs and compare them with typical face recognition models with similar protocols. To our knwoledge, this paper is the first study on the application of MLLMs for heterogeneous face recognition.

3. Evaluating MLLMs for Heterogeneous Face Recognition

In this section, we describe how to use Multimodal Large Language Models (MLLMs) for heterogeneous face recognition (HFR) and our evaluation pipeline for existing models under standardized biometric protocols.

3.1. Heterogeneous Face Recognition

Heterogeneous face recognition involves determining whether two face images belong to the same identity when the images are acquired using different sensing modalities. Let (x_e, x_p) denote an enrollment-probe image pair, where x_e and x_p may originate from different modalities (e.g., VIS-NIR, VIS-SWIR, VIS-THERMAL). The task is to output a similarity score or decision indicating whether the given pair of images corresponds to the same identity.

3.2. Model Inference and Prompting

We consider each MLLM in a zero-shot setting using a fixed prompt template designed to elicit identity comparison between two face images. As a system prompt, we first describe the task for the MLLM: *"You are an AI assistant specialised in heterogeneous face recognition."* Then, the user prompt instructs the model to assess whether the two images (from two different modalities) depict the same individual, without providing any identity labels:

Prompt **User**

I give you two face images taken under `modality_1` and `modality_2`. On a scale from 0 to 100, how likely (as a single number) are these two faces of the same person? Only output a single number (no other text)."

As we discuss later in our experiment setup, `modality_1` and `modality_2` can be visual, NIR, SWIR, or thermal.

Based on the prompt, the output of MLLM is expected to be a single number as similarity score of two images. This score is then used to construct genuine and impostor score distributions following standard biometric evaluation practices. As baselines, we use a classical face recognition model and a model trained for HFR setting. We benchmark these baselines with the same protocols by extracting embeddings and calculating cosine similarity of embeddings.

3.3. Performance Metrics

We evaluate the recognition performance is evaluated using standard biometric metrics. We report the Equal Error Rate (EER), defined as the operating point where the False Accept Rate (FAR) equals the False Reject Rate (FRR). In addition, we report the True Accept Rate (TAR) at a fixed FAR of 1%, which reflects performance under a security-constrained operating condition.

While we expect a similarity score for each pair of images, not all models successfully process every image pair. We explicitly account for such cases by reporting the *Acquire Rate (AR)*, defined as the proportion of image pairs for which a valid score is produced. Image pairs for which a model fails to return a usable output are counted as failures to acquire and excluded from subsequent matching performance metrics.

4. Experiments

4.1. Experimental Setup

4.1.1 Dataset

In our experiments, we use three different datasets and evaluate HFR with color (RGB), near-infrared (NIR), short-wave infrared (SWIR), and thermal images. For each dataset, we consider 1,000 random positive pairs (from different modalities) and 1,000 random negative pairs.

MCXFace Dataset: The MCXFace dataset [14] consists of multi-channel image samples for 51 subjects. For each subject color (RGB), thermal, near-infrared (850 nm), short-wave infrared (1300 nm), depth, stereo depth, and depth estimated from RGB images are collected in differ-

ent channels under three different sessions and various illumination conditions. We use color (RGB), thermal, near-infrared images of this dataset for evaluation of HFR under VIS-NIR and VIS-THERMAL conditions.

CASIA NIR-VIS 2.0 Dataset: The CASIA NIR-VIS 2.0 Face dataset [27] includes images of 725 individuals captured under real-world, unconstrained acquisition conditions with variations in pose, expression, and illumination for both visible spectrum and near-infrared (NIR). Each subject in the dataset is captured in both NIR and VIS modalities across multiple imaging sessions (1-22 visible spectrum photos and 5-50 NIR photos.)

Polathermal Dataset: The Polathermal dataset [18], collected by the U.S. Army Research Laboratory (ARL), consists of both polarimetric long-wave infrared (LWIR) imagery with Stokes parameter representations and color images. The dataset includes 60 subjects acquired under varying distances. In our experiments, we use the thermal and color images from this dataset. We use 25 identities for the training set (of baseline HFR) and the remaining 35 identities for the test set.

4.1.2 Pretrained MLLMs

We consider multiple open-source MLLMs with different sizes and structures, and evaluate their performance for HFR:

Gemma 3: Gemma 3 [46] is a MLLM developed by Google from the Gemma family [47] of lightweight open large language models, ranging in scale from 1 to 27 billion parameters, that processes both text and images and generates text output. These models are able to handle long context and can support up to a 128 k-token context window. This is achieved by increasing the ratio of local to global attention layers, while keeping the span on local attention short. The Gemma 3 models are trained with distillation and achieve superior performance to Gemma 2 for both pre-trained and instruction finetuned versions. We use Gemma-3-4B-it and Gemma-3-27B-it models in our experiments.

LLaVA: LLaVA [31] is a vision-language model that extends LLMs with visual understanding capabilities through multi-stage training which combines vision encoders with instruction-tuned language models. It leverages image-text instruction data and alignment techniques to enable processing visual input along with textual input, achieving significant performance on a wide range of multimodal benchmarks while being lightweight and efficient. LLaVA-1.5 achieved significant improvement in visual question answering, image description, and general-purpose multimodal interaction, making it a powerful model for downstream vision-language applications. We use LLaVA-1.5-7B model in our experiments.

Phi-4-Multimodal: Phi-4-Multimodal [1] is a compact yet high-performance MLLM that extends the Phi family with integrated vision understanding capabilities. The model is designed with an emphasis on efficiency and high-quality training data. Phi-4-Multimodal model combines a strong small language backbone with aligned visual encoders to support tasks, such as image captioning, visual question answering, document and chart understanding, and multimodal reasoning. Because of its relatively small parameter scale, Phi-4-Multimodal is particularly suitable for resource-constrained environments and edge deployment scenarios.

Mistral Mistral-3.1 is an open-source MLLM developed by Mistral AI that has significant performance with 24 billion parameters. It builds on its predecessor by integrating strong text and vision understanding, supporting a long context window of up to 128 k tokens, and delivering competitive performance across instruction following, conversational assistance, multilingual tasks, code generation, and long-document reasoning. Mistral-3.1 achieved fast inference speeds and can be deployed on consumer-grade hardware while maintaining high capability across diverse generative and analytic workloads. We use Mistral-3.1-24B model[1] in our experiments.

Aya-Vision: The Aya-Vision family [6] comprises open-weights MLLMs developed by Cohere that integrate advanced visual understanding with multilingual text capabilities across 23 languages. Built by combining a multilingual language backbone with a state-of-the-art vision encoder, Aya-Vision models excel at several vision-language tasks, such as image captioning, optical character recognition, visual reasoning, and visual question answering. These models achieve competitive or leading performance on multilingual multimodal benchmarks. We use Aya-Vision-8B and Aya-Vision-32B models in our experiments.

InternVL2: InternVL2 [5] is an open-source MLLM that adopts a unified ViT-MLP-LLM architecture, combining a vision transformer (InternViT), a lightweight projection module, and a LLM to enable joint visual and textual reasoning. The model is trained on diverse multimodal data, including images, documents, charts, OCR

[1] Available at https://huggingface.co/mistralai/Mistral-Small-3.1-24B-Instruct-2503

data, videos, and long-form text. InternVL2 supports multi-image and video inputs with dynamic resolution processing and demonstrates strong performance across tasks such as visual question answering, document understanding, OCR, and scientific reasoning.

InternVL3: InternVL3 [54] is a series of open-source MLLM that unifies vision and language capabilities through a native multimodal pre-training paradigm, jointly learning from both multimodal and pure-text data in a single stage. It adopts a ViT-MLP-LLM architecture with a new variable visual position encoding (V2PE) to handle extended multimodal contexts effectively. InternVL3 models achieved competitive performance on a wide range of multimodal reasoning, visual comprehension, and language tasks. We use InternVL3-9B in our experiments.

Qwen3-VL Qwen3-VL [3] is a family of open-weight MLLMs developed by Alibaba that has multimodal understanding across text, images, and video. It combines a vision encoder with a LLM decoder, enabling understanding and reasoning over visual and textual inputs with a native long context of up to 256 K tokens (expandable toward 1 M) for processing long documents and videos. Qwen3-VL achieved competitive performance across diverse benchmarks, such as visual question answering, OCR, spatial and 2D/3D reasoning, etc. We use Qwen3-VL-4B-Instruct and Qwen3-VL-8B-Instruct in our experiments.

4.1.3 Baselines

As baselines in our experiments, we compare the performance of MLLMs with a heterogeneous face recognition model (xEdgeFace) for different modalities and also a classical face recognition model (EdgeFace). These two baseline models have similar network structure and are different in training.

EdgeFace: EdgeFace [11] is a state-of-the-art lightweight face recognition model based on CNN-Transformer backbone. This model is trained on RGB images from the WebFace dataset [56], and we do not further finetune it for other modalities in our HFR experiments. Our goal is to use this model as a baseline to see how a model trained for visual (RGB) images perform for HFR (for comparing with MLLMs in zero-shot setup). While EdgeFace is a lightweight face recognition model, it is amongst top-performing models on different face recognition benchmarks.

xEdgeFace: xEdgeFace [13] is a contrastive self-distillation framework for adapting a pretrained face recognition backbone to heterogeneous face recognition. The method retains the original EdgeFace [11] architecture and introduces cross-modal alignment together with teacher–student supervision to enable effective adaptation while preserving performance on the source (RGB) face recognition task. As a result, the model acquires robust heterogeneous recognition capability without catastrophic forgetting. Despite its lightweight design, xEdgeFace achieves performance comparable to or surpassing state-of-the-art methods on multiple HFR benchmarks.

4.1.4 Implementation Details

We ran our experiments on a system equipped with NVIDIA H100. We used the vLLM repository[2] to implement our evaluation of MLLM for HFR. The source code of our experiments is publicly available[3].

4.2. Experimental Results

Performance on MCXFace Dataset: Table 1 reports the performance of multiple open-source MLLMs for HFR under VIS-NIR, VIS-SWIR, and VIS-THERMAL modalities from the MCXFace dataset. The table also compares the performance of MLLMs with pretrained face recognition (EdgeFace) and HFR (xEdgeFace) models. For the performance of each model, we report Acquire Rate (AR), Equal Error Rate (EER), and True Accept Rate (TAR) at FAR of 1%. While the majority of MLLMs have AR of 100%, in some cases (for Mistral-3.1 and Aya-Vision), the model was not able to provide similarity score for some samples.

For all modalities in Table 1, we observe that the pretrained HFR model (xEdgeFace) achieves better results than MLLMs. In most tasks, the pretrained face recognition (EdgeFace) also has a better performance than MLLMs, except for VIS-THERMAL, where Qwen3-VL-4B-Instruct gets a better EER value than EdgeFace. However, for VIS-THERMAL, EdgeFace has a better TAR value. Among different modalities, VIS-NIR appears to be the easiest for MLLMs and VIS-THERMAL is the hardest task. This can be due to the fact that MLLMs have been mainly trained on visual (RGB) images. While NIR images are similar to gray images, thermal images look substantially different. Meanwhile, we should note that VIS-THERMAL is also the most difficult modality for pretrained face recognition (EdgeFace) and HFR (xEdgeFace) models.

In the remaining experiments, we focus on top-performing MLLMs in Table 1 and further evaluate their performance for HFR.

Performance on CASIA Dataset: Table 2 compares the performance of MLLMs for VIS-NIR modality with baselines (xEdgeFace and EdgeFace) on the CASIA dataset.

[2] https://github.com/vllm-project/vllm
[3] https://gitlab.idiap.ch/biometric/code.hfr_mllm

Table 1. Heterogeneous Face Recognition (HFR) performance across modalities in the MCXFace dataset. AR denotes Acquire Rate. TAR is reported at FAR=1%.

Model	Modality								
	VIS–NIR			VIS–SWIR			VIS–THERMAL		
	AR↑	EER↓	TAR↑	AR↑	EER↓	TAR↑	AR↑	EER↓	TAR↑
Open source MLLMs									
Gemma-3-4B-it	100%	17.40%	0.00%	100%	32.00%	0.00%	100%	52.00%	0.00%
Gemma-3-27B-it	100%	5.80%	35.60%	100%	21.00%	**14.70%**	100%	38.70%	0.80%
LLaVA-1.5-7B-hf	100%	50.20%	0.20%	100%	52.36%	0.40%	100%	47.95%	0.40%
Phi-4-Multimodal-Instruct	73%	50.14%	1.76%	76%	49.54%	0.26%	68%	49.85%	1.30%
Mistral-3.1-24B-Instruct	94%	36.88%	1.79%	93%	40.81%	2.89%	89%	45.42%	0.11%
Aya-Vision-8B	100%	29.05%	0.50%	99%	58.69%	0.10%	100%	50.45%	0.10%
Aya-Vision-32B	90%	38.60%	0.22%	91%	47.29%	2.09%	89%	50.74%	1.00%
InternVL2-4B	96%	42.71%	0.00%	93%	46.20%	1.72%	96%	48.12%	2.08%
InternVL3-9B	100%	7.93%	21.94%	100%	25.20%	5.53%	100%	43.22%	**9.44%**
Qwen3-VL-4B-Instruct	100%	3.80%	84.20%	100%	52.40%	5.80%	100%	**16.30%**	9.00%
Qwen3-VL-8B-Instruct	100%	**2.90%**	**85.70%**	100%	**12.10%**	13.00%	100%	28.80%	4.00%
Face Recognition Models									
xEdgeFace	100%	0.00%	100%	100%	0.20%	99.80%	100%	5.80%	82.00%
EdgeFace	100%	0.00%	100%	100%	0.70%	99.50%	100%	22.50%	30.30%

Table 2. VIS–NIR heterogeneous face recognition results on the CASIA datset.

Model	Acquire Rate	EER	TAR@FAR=1%
Open source MLLMs			
Qwen3-VL-8B-Instruct	100%	2.90%	35.10%
InternVL3-9B	100%	16.82%	23.40%
Gemma-3-27B-it	100%	21.56%	0.00%
Face Recognition Models			
xEdgeFace	100%	0.00%	100%
EdgeFace	100%	0.10%	99.84%

Table 3. VIS–THERMAL Heterogeneous Face Recognition results on the Polarimetric dataset.

Model	Acquire Rate	EER	TAR@FAR=1%
Open source MLLMs			
Qwen3-VL-8B-Instruct	100%	12.90%	25.40%
InternVL3-9B	100%	27.56%	13.86%
Gemma-3-4B-it	100%	35.70%	0.00%
Face Recognition Models			
xEdgeFace	100%	1.60%	97.70%
EdgeFace	100%	22.80%	33.60%

Similar to VIS-NIR evaluation on the MCXFace dataset (Table 1), we can see that MLLMs have inferior performance in terms of EER and TAR compared to baselines (xEdgeFace and EdgeFace) on the CASIA dataset. Comparing the values with Table 1, Qwen3-VL has same EER but much lower TAR value. Gemma-3 also has inferior performance in terms of EER and TAR. In contrast, InternVL3 has worse EER but slightly better TAR value that was achieved on the MCXFace dataset.

Performance on Polarimetric Dataset: Table 3 also reports the performance of MLLMs for VIS-THERMAL heterogeneous face recognition on the Polarimetric dataset and compares with baselines (xEdgeFace and EdgeFace). As can be seen, thermal images are difficult for MLLMs to match, which makes VIS-THERMAL heterogeneous face recognition a challenging task. Compared to Table 1, all MLLMs achieved better EER values and better TAR (except for Gemma-3 which has slightly lower TAR). Similar to Table 1, the pretrained HFR model (xEdgeFace) has the best performance. While Qwen3-VL achieves better EER than EdgeFace, still EdgeFace performs better in terms of TAR.

In-Context Learning: As another experiment, we evaluate MLLMs under in-context learning setup, in which we provide example of images from both modalities before asking the MLLM to provide similarity score for test images. The new system prompt is as follows: *"You are an AI assistant specialised in heterogeneous face recognition. I give you two face images taken under* `modality_1` *and* `modality_2`*. On a scale from 0 to 100, how likely (as a single number) are these two faces of the same person? Only output a single number (no other text)."* Then we provide examples of images from different modalities but from the same subject with the following prompt: *"Here is an*

Table 4. In-context learning using Qwen3-VL-8B-Instruct across modalities in the MCXFace dataset. AR denotes Acquire Rate. TAR is reported at FAR=1%.

No. Examples	Modality								
	VIS–NIR			VIS–SWIR			VIS–THERMAL		
	AR↑	EER↓	TAR↑	AR↑	EER↓	TAR↑	AR↑	EER↓	TAR↑
0 (Zero-shot)	100%	**2.90%**	**85.70%**	100%	**12.10%**	**13.00%**	100%	**28.80%**	4.00%
1 positive + 1 negative	100%	8.80%	72.00%	100%	28.20%	4.50%	100%	**28.80%**	**13.20%**

*example of two images of the **same** person taken under* `modality_1` *and* `modality_2`*. You need to output a high similarity score for such pairs."* Similarly, we provide examples of images from different modalities and from different subjects with the following prompt; *"Here is an example of two images of **different** people taken under* `modality_1` *and* `modality_2`*. You need to output a low similarity score for such pairs."* Table 4 compares the performance of Qwen3-VL-8B-Instruct in zero-shot setting versus in-context learning with one positive pair and one negative pair examples. While in-context learning provides more information through example, it has not improved the performance of the model. This can indicate the difficulty of HFR which requires matching faces based on fine-grained details in two images from different modalities to distinguish the identity in the open-set evaluation.

MLLM Reasoning: As another study, we investigate the reasoning process of MLLMs for HFR, when requested to provide explanations. To this end, we ask the model provide explanation using the following prompt: *"Analyze the facial features and explain your reasoning about whether these two faces belong to the same person."* We focus on Qwen3-VL-8B-Instruct on VIS-NIR scenario. pairs from the MCXFace dataset, one genuine pair (same identity) and one impostor pair (different identities), both of which were correctly classified by the model. Table 5 presents two correctly classified VIS-NIR pairs from the MCXFace dataset alongside the full output of Qwen3-VL-8B-Instruct. The green text highlights correct and reliable observations about facial geometry and cross-spectral imaging physics in the MLLM output, while blue highlights reasoning that relies on mutable appearance attributes (e.g., accessories, facial hair) which are not reliable biometric cues.

In the genuine pair (score: 98/100), the model correctly identifies structural consistency across modalities (forehead width, eye shape, nose bridge, and jawline) and explicitly interprets the reduced contrast and texture in the NIR image as a sensor effect rather than an identity mismatch. This demonstrates a meaningful understanding of cross-spectral imaging physics. In the impostor pair (score: 0/100), the model correctly detects geometric mismatches in nose shape, jawline, eye shape, and brow ridge that persist despite modality differences, and explicitly accounts for cross-modal appearance changes when forming its decision. However, part of the reasoning also refers to glasses, mustache, and beard style as distinguishing cues, which are mutable and unreliable for biometric recognition.

These examples suggest that MLLMs can produce partially correct and interpretable cross-modal reasoning. However, using unreliable cues along with the overall EER values in Tables 1-4 shows that MLLM reasoning does not guarantee reliable recognition performance.

5. Conclusion

In this paper, we explored the application of MLLMs for heterogeneous face recognition. Heterogeneous face recognition requires comparison of face images from two different modalities. Multimodal large language models are often trained on general datasets, including RGB images. Although MLLMs have achieved remarkable performance on different downstream computer vision tasks, they are primarily trained on general-purpose data dominated by RGB imagery. Our experimental results show that, despite their recent advances, current MLLMs struggle to match information from different image modalities for heterogeneous face recognition, including VIS–NIR, VIS–SWIR, and VIS–THERMAL scenarios. In comparison to classical face recognition systems specifically developed for face biometric, MLLMs have substantial performance gaps, particularly under challenging cross-spectral conditions. We also investigated in-context learning by showing MLLM example of positive and negative pairs of images from different modalities. However, our findings indicate such examples could not be sufficient for MLLMs to enhance their understanding of given images and distinguish two face images from different modalities. While MLLM can also provide interpretable cross-modal reasoning, they may use unreliable cues which may affect the recognition performance. Overall, our study highlights the current limitations of MLLMs for heterogeneous face recognition and shows the importance of modality-aware training and also rigorous evaluation of the models for biometrics. While MLLMs hold promise as general-purpose multimodal models, significant advances are required before they can be reliably applied to sensitive biometric recognition tasks.

Table 5. Two example of VIS-NIR pairs from the MCXFace dataset, correctly classified by Qwen3-VL-8B-Instruct. Each row shows the input images and the model's full reasoning. Green highlights indicate correct and reliable observations about facial geometry and cross-spectral imaging physics; blue highlights indicate reasoning that relies on mutable appearance attributes (glasses, beard, hairstyle) which are not dependable biometric cues.

VIS	NIR	MLLM Output
		**Genuine pair (same identity)** **Score: 98/100** ✓ Upon detailed analysis of the two provided facial images - one captured in Visible (VIS) light and the other in Near-Infrared (NIR) - I observe the following: Facial Structure and Landmarks: - The overall facial structure, including the width of the forehead, the shape of the eyes, the nose bridge and tip, the mouth, and jawline, is consistent across both modalities. The proportions of the face, such as the eye-to-nose distance and the nasal width, appear nearly identical. - Key facial landmarks (e.g., the inner and outer corners of the eyes, the tip of the nose, the corners of the mouth) align remarkably well in both images, indicating a strong match in underlying anatomical structure. Cross-Modality Differences: - As expected, the NIR image exhibits lower contrast and color saturation compared to the VIS image, which is common due to the different spectral sensitivities of the sensors. NIR light penetrates deeper and reflects off skin differently. - The VIS image shows skin texture, color variations, and fine details (such as stubble, pores, and subtle skin tone differences) that are absent or muted in the NIR image. - The NIR image may show slightly different albedo (reflectivity) and shadow patterns caused by varying surface properties and light absorption in the infrared spectrum. Conclusion: The preserved geometric relationships and consistent landmark positions across both modalities strongly suggest that these two images are of the same individual. The cross-modality differences are consistent with known physical properties of VIS and NIR imaging, rather than indicating a mismatch. There are no discernible anatomical discrepancies (e.g., asymmetry, mismatched landmarks) to suggest different subjects.
		**Impostor pair (different identities)** **Score: 0/100** ✓ Upon analyzing the two provided face images – one appearing to be in visible light (VIS) and the other in near-infrared (NIR) – I find significant discrepancies that strongly suggest they depict two different individuals. First, examining the facial structure and landmarks: - The first image (VIS) shows a man with a receding hairline, glasses, and a well-defined mustache and goatee. His facial bone structure appears relatively slender. - The second image (NIR) shows a man with a full, dark beard, thick, wavy hair, and a much more robust, wider facial structure with prominent cheekbones. - The shape of the nose, jawline, eye shape, and brow ridge are fundamentally different between the two subjects. - In the VIS image, the eyebrows are relatively sparse and light, whereas in the NIR image, they are well-defined and dark, consistent with bearded facial hair. Second, considering cross-modality differences: - NIR imaging often enhances certain features (like skin texture and underlying vasculature) while reducing the impact of surface details that are visible under normal lighting, such as hair color and skin tone. - However, even after accounting for these modality differences, the fundamental facial geometry – the relative position of the eyes, nose, and mouth, the width of the face, and the contours of the jawline – remains distinct. The presence of a beard in the NIR image is a key differentiator, as it is a major and permanent facial feature. While it is technically possible for two people to have similar facial structures and be photographed under different lighting conditions, the spatial and geometric differences observed here are too pronounced to be dismissed as a single person under dual modality. The difference in facial hair (glasses vs. beard), facial width, and overall bone structure indicate two separate individuals.

Ethics Statement

This work investigates the application of MLLMs for heterogeneous face recognition, which is a biometric task with clear ethical, legal, and societal implications. Our study is conducted with the goal of understanding potentials and limitations of MLLMs when used for HFR. All our experiments are performed on existing publicly available heterogeneous face datasets collected under established research protocols. We also used open-source MLLMs in our study and excluded commercial models, ensuring that biometric data in our evaluation is not transmitted to a third party. Overall, this work contributes to the responsible development of biometric systems by providing an assessment of MLLMs for HFR and by highlighting the necessity of rigorous evaluation before considering their use in sensitive identity-related applications.

Acknowledgment

This work was funded by the European Union project CarMen (Grant Agreement No. 101168325).

References

[1] M. Abdin, J. Aneja, H. Behl, S. Bubeck, R. Eldan, S. Gunasekar, M. Harrison, R. J. Hewett, M. Javaheripi, P. Kauffmann, et al. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*, 2024. 4

[2] M. Awais, M. Naseer, S. Khan, R. M. Anwer, H. Cholakkal, M. Shah, M.-H. Yang, and F. S. Khan. Foundation models defining a new era in vision: a survey and outlook. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2025. 1

[3] S. Bai, Y. Cai, R. Chen, K. Chen, X. Chen, et al. Qwen3-vl technical report. *arXiv preprint arXiv:2511.21631*, 2025. 5

[4] U. Cheema, M. Ahmad, D. Han, and S. Moon. Heterogeneous visible-thermal and visible-infrared face recognition using cross-modality discriminator network and unit-class loss. *Computational Intelligence and Neuroscience*, 2022, 2022. 2

[5] Z. Chen, J. Wu, W. Wang, W. Su, G. Chen, S. Xing, M. Zhong, Q. Zhang, X. Zhu, L. Lu, et al. Internvl: Scaling up vision foundation models and aligning for generic visual-linguistic tasks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 24185–24198, 2024. 4

[6] S. Dash, Y. Nan, J. Dang, A. Ahmadian, S. Singh, M. Smith, B. Venkitesh, V. Shmyhlo, V. Aryabumi, W. Beller-Morales, et al. Aya vision: Advancing the frontier of multilingual multimodality. *arXiv preprint arXiv:2505.08751*, 2025. 4

[7] T. de Freitas Pereira, A. Anjos, and S. Marcel. Heterogeneous face recognition using domain specific units. *IEEE Transactions on Information Forensics and Security*, 14(7):1803–1816, 2018. 2

[8] J. Deng, J. Guo, N. Xue, and S. Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 4690–4699, 2019. 1

[9] P. Farmanifard and A. Ross. Chatgpt meets iris biometrics. In *2024 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–10. IEEE, 2024. 2

[10] C. Fu, X. Wu, Y. Hu, H. Huang, and R. He. DVG-face: Dual variational generation for heterogeneous face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021. 2

[11] A. George, C. Ecabert, H. O. Shahreza, K. Kotwal, and S. Marcel. EdgeFace: Efficient face recognition model for edge devices. *IEEE Transactions on Biometrics, Behavior, and Identity Science*, 2024. 1, 2, 5

[12] A. George and S. Marcel. Bridging the gap between visible and thermal face recognition. In *Proceedings of the IEEE/CVF Conference*, 2023. 2

[13] A. George and S. Marcel. xedgeface: Efficient cross-spectral face recognition for edge devices. pages 1–10, 2025. 2, 5

[14] A. George, A. Mohammadi, and S. Marcel. Prepended domain transformer: Heterogeneous face recognition without bells and whistles. *IEEE Transactions on Information Forensics and Security*, 18:133–146, 2022. 3

[15] A. Hassanpour, Y. Kowsari, H. O. Shahreza, B. Yang, and S. Marcel. Chatgpt and biometrics: an assessment of face recognition, gender detection, and age estimation capabilities. In *2024 IEEE International Conference on Image Processing (ICIP)*, pages 3224–3229. IEEE, 2024. 1, 2

[16] R. He, X. Wu, Z. Sun, and T. Tan. Learning invariant deep representation for nir-vis face recognition. In *Thirty-First AAAI Conference on Artificial Intelligence*, 2017. 2

[17] R. He, X. Wu, Z. Sun, and T. Tan. Wasserstein cnn: Learning invariant features for nir-vis face recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 41(7):1761–1773, 2018. 2

[18] S. Hu, N. J. Short, B. S. Riggan, C. Gordon, K. P. Gurton, M. Thielke, P. Gurram, and A. L. Chan. A polarimetric thermal database for face recognition research. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 119–126, 2016. 4

[19] A. Hurst, A. Lerer, A. P. Goucher, A. Perelman, A. Ramesh, A. Clark, A. Ostrow, A. Welihinda, A. Hayes, A. Radford, et al. Gpt-4o system card. *arXiv preprint arXiv:2410.21276*, 2024. 2

[20] S. Jia, R. Lyu, K. Zhao, Y. Chen, Z. Yan, Y. Ju, C. Hu, X. Li, B. Wu, and S. 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*, pages 4324–4333, 2024. 2

[21] M. Kan, S. Shan, H. Zhang, S. Lao, and X. Chen. Multi-view discriminant analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 38(1):188–194, 2015. 2

[22] B. Klare, Z. Li, and A. K. Jain. Matching forensic sketches to mug shot photos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(3):639–646, 2010. 2

[23] B. F. Klare and A. K. Jain. Heterogeneous face recognition using kernel prototype similarities. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 35(6):1410–1422, 2012. 2

[24] A. Komaty, H. O. Shahreza, A. George, and S. Marcel. Exploring chatgpt for face presentation attack detection in zero and few-shot in-context learning. In *Proceedings of the Winter Conference on Applications of Computer Vision*, pages 834–843, 2025. 2

[25] Z. Lei and S. Z. Li. Coupled spectral regression for matching heterogeneous faces. In *2009 IEEE Conference on Computer Vision and Pattern Recognition*, pages 1123–1128. IEEE, 2009. 2

[26] Z. Lei, S. Liao, A. K. Jain, and S. Z. Li. Coupled discriminant analysis for heterogeneous face recognition. *IEEE Transactions on Information Forensics and Security*, 7(6):1707–1716, 2012. 2

[27] S. Li, D. Yi, Z. Lei, and S. Liao. The casia nir-vis 2.0 face database. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 348–353, 2013. 4

[28] S. Liao, M. W. K. Law, and A. C. S. Chung. Learning multiscale block local binary patterns for face recognition. In *Advances in Biometrics*, pages 828–837. Springer, 2007. 2

[29] S. Liao, D. Yi, Z. Lei, R. Qin, and S. Z. Li. Heterogeneous face recognition from local structures of normalized appearance. In *International Conference on Biometrics*, pages 209–218. Springer, 2009. 2

[30] D. Lin and X. Tang. Inter-modality face recognition. In *European Conference on Computer Vision*, pages 13–26. Springer, 2006. 2

[31] H. Liu, C. Li, Y. Li, and Y. J. Lee. Improved baselines with visual instruction tuning. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 26296–26306, 2024. 4

[32] Q. Liu, X. Tang, H. Jin, H. Lu, and S. Ma. A nonlinear approach for face sketch synthesis and recognition. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 1005–1010. IEEE, 2005. 2

[33] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004. 2

[34] K. Narayan, V. VS, and V. M. Patel. Facexbench: Evaluating multimodal llms on face understanding. *arXiv preprint arXiv:2501.10360*, 2025. 1, 3

[35] T. Ojala, M. Pietikainen, and T. Maenpaa. Multiresolution gray-scale and rotation invariant texture classification with local binary patterns. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(7):971–987, 2002. 2

[36] U. Ozturk, H. O. Shahreza, and S. Marcel. Demographic fairness in multimodal llms: A benchmark of gender and ethnicity bias in face verification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 1184–1193, 2026. 3

[37] L. Qin et al. Face-human-bench: A comprehensive benchmark of face and human understanding for multi-modal assistants. *arXiv preprint arXiv:2501.01243*, 2025. 3

[38] H. O. Shahreza and S. Marcel. Benchmarking multimodal large language models for face recognition. *arXiv preprint arXiv:2510.14866*, 2025. 3

[39] H. O. Shahreza and S. Marcel. Facellm: A multimodal large language model for face understanding. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 3677–3687, 2025. 1

[40] H. O. Shahreza and S. Marcel. Foundation models and biometrics: A survey and outlook. *IEEE Transactions on Information Forensics and Security*, 2025. 1, 2

[41] A. Sharma and D. W. Jacobs. Bypassing synthesis: Pls for face recognition with pose, low-resolution and sketch. In *CVPR 2011*, pages 593–600. IEEE, 2011. 2

[42] Y. Shi, Y. Gao, Y. Lai, H. Wang, J. Feng, L. He, J. Wan, C. Chen, Z. Yu, and X. Cao. Shield: An evaluation benchmark for face spoofing and forgery detection with multimodal large language models. *arXiv preprint arXiv:2402.04178*, 2024. 2

[43] R. Sony, P. Farmanifard, H. Alzwairy, N. Shukla, and A. Ross. Benchmarking foundation models for zero-shot biometric tasks. *arXiv preprint arXiv:2505.24214*, 2025. 1

[44] R. Sony, P. Farmanifard, A. Ross, and A. K. Jain. Foundation versus domain-specific models: Performance comparison, fusion, and explainability in face recognition. *arXiv preprint arXiv:2507.03541*, 2025. 3

[45] X. Tang and X. Wang. Face sketch synthesis and recognition. In *Proceedings Ninth IEEE International Conference on Computer Vision*, pages 687–694. IEEE, 2003. 2

[46] G. Team, A. Kamath, J. Ferret, S. Pathak, N. Vieillard, R. Merhej, S. Perrin, T. Matejovicova, A. Ramé, M. Rivière, et al. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*, 2025. 4

[47] G. Team, T. Mesnard, C. Hardin, R. Dadashi, S. Bhupatiraju, S. Pathak, L. Sifre, M. Rivière, M. S. Kale, J. Love, et al. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*, 2024. 4

[48] X. Wang, X. Ma, X. Hou, M. Ding, Y. Li, J. Chen, W. Chen, X. Peng, and L. Shen. Facebench: A multi-view multi-level facial attribute vqa dataset for benchmarking face perception mllms. *arXiv preprint arXiv:2503.21457*, 2025. 3

[49] X. Wang and X. Tang. Face photo-sketch synthesis and recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 31(11):1955–1967, 2008. 2

[50] D. Yi, R. Liu, R. Chu, Z. Lei, and S. Z. Li. Face matching between near infrared and visible light images. In *International Conference on Biometrics*, pages 523–530. Springer, 2007. 2

[51] H. Zhang, V. M. Patel, B. S. Riggan, and S. Hu. Generative adversarial network-based synthesis of visible faces from polarimetric thermal faces. In *2017 IEEE International Joint Conference on Biometrics (IJCB)*, pages 100–107. IEEE, 2017. 2

[52] J. Zhang, J. Huang, S. Jin, and S. Lu. Vision-language models for vision tasks: A survey. *IEEE transactions on pattern analysis and machine intelligence*, 46(8):5625–5644, 2024. 1

[53] W. Zhang, X. Wang, and X. Tang. Coupled information-theoretic encoding for face photo-sketch recognition. In *CVPR 2011*, pages 513–520. IEEE, 2011. 2

[54] J. Zhu, W. Wang, Z. Chen, Z. Liu, S. Ye, L. Gu, H. Tian, Y. Duan, W. Su, J. Shao, et al. Internvl3: Exploring advanced training and test-time recipes for open-source multimodal models. *arXiv preprint arXiv:2504.10479*, 2025. 5

[55] J.-Y. Zhu, T. Park, P. Isola, and A. A. Efros. Unpaired image-to-image translation using cycle-consistent adversarial networks. *Proceedings of the IEEE International Conference on Computer Vision*, pages 2223–2232, 2017. 2

[56] Z. Zhu, G. Huang, J. Deng, Y. Ye, J. Huang, X. Chen, J. Zhu, T. Yang, J. Lu, D. Du, et al. Webface260m: A benchmark unveiling the power of million-scale deep face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10492–10502, 2021. 5