Zero-Shot Biometric Capabilities of the Gemma 3 Vision-Language Model Family

Sergio Picascia, Elisabetta Rocchetti

1 Introduction

Vision-Language Models (VLMs) have emerged as powerful tools capable of jointly processing and understanding visual and textual information, offering capabilities in multimodal data interpretation. This report investigates the application and performance of these AI systems, specifically the recently released Gemma 3 family of models [Gemma Team, 2025], in the context of various biometric tasks.

Biometrics traditionally relies on specialized algorithms tailored for specific modalities. However, the advent of general-purpose VLMs presents an opportunity to explore their potential for zero-shot or few-shot learning in biometrics, potentially reducing the need for extensive task-specific training and offering enhanced interpretability. This study aims to provide a comprehensive evaluation of VLMs across a diverse set of biometric challenges, including face, iris, and fingerprint verification, as well as soft biometric tasks like age and gender estimation, and face attribute classification. By benchmarking their performance against established datasets and state-of-the-art (SOTA) specialized systems, we seek to understand the current capabilities, limitations, and future prospects of VLMs in contributing to the evolving landscape of biometric identification and analysis.

This report begins with an overview of VLMs and foundation models, details the methodology for evaluation, presents experimental results across multiple biometric tasks, and concludes with a discussion of the findings and their implications.

2 Preliminaries

2.1 Vision-language models

VLMs are advanced multimodal AI systems designed to process and understand both visual (images or videos) and textual data simultaneously. They integrate computer vision and natural language processing (NLP) capabilities, enabling machines to interpret, relate, and generate content that involves both modalities. VLMs typically consist of two main components:

- Vision Encoder: Processes images or videos to extract meaningful visual features such as shapes, colors, textures, and spatial relationships. Modern VLMs such as CLIP [Radford et al., 2021] often employ vision transformers (ViTs) [Dosovitskiy et al., 2021], which treat images as sequences of patches and use self-attention mechanisms to capture global context.
- Text Encoder: Converts textual input into semantic embeddings that capture the meaning and context of words and phrases. This encoder is usually based on transformer architectures like BERT [Devlin et al., 2019] or GPT [Brown et al., 2020], which use self-attention to understand language context and generate embeddings.

These encoders map their respective inputs into a shared embedding space, allowing the model to align and correlate visual and textual information effectively. VLMs work by encoding images and texts into embeddings and then aligning these embeddings so that related image-text pairs are close in the latent space, while unrelated pairs are far apart. This alignment is often achieved through contrastive learning, where the model is trained to push matching image-text pairs together and non-matching pairs apart. Once trained, the model can perform various tasks such as image captioning, visual question-answering, image-text retrieval.

2.2 Foundation models in biometrics

Given their remarkable zero- and few-shot performances, VLMs have been tested for biometrics tasks including biometric recognition and soft-biometric detection. We report here a selection of the most recent works about zero-shot classification with VLMs discussed in the survey Shahreza and Marcel [2025].

Hassanpour et al. [2024] specifically assessed the performance of GPT-4V [OpenAI, 2023] on tasks such as face recognition, gender classification, and age estimation. Their findings indicate that GPT-4V is capable of

distinguishing between facial identities with notable accuracy, achieves strong performance in gender detection, and performs reasonably well in age estimation.

Deandres-Tame et al. [2024] explored the capabilities of GPT-4V in face verification and soft biometric attribute estimation. The authors emphasize GPT-4V's potential to enhance the transparency of automated decisions in human-centered scenarios by providing textual explanations for its outputs. Their study suggests that LLMs like GPT-4V exhibit competitive performance when compared to dedicated face verification and soft biometric systems, including ArcFace [Deng et al., 2022], AdaFace [Kim et al., 2022], and FairFace [Karkkainen and Joo, 2021].

Farmanifard and Ross [2024] investigated the application of GPT-4V's multimodal capabilities to iris recognition—a less commonly studied yet crucial biometric task. Their experiments demonstrate GPT-4V's precision and adaptability in analyzing iris features, including the detection of subtle factors such as the presence of makeup. Compared to the state-of-the-art VeriEye matcher [Neurotechnology, 2025], GPT-4V shows superior performance. Additionally, the study reports that GPT-4V outperforms Gemini [Gemini Team, 2025] in both recognition accuracy and user experience.

Li et al. [2024] introduce CLIPER, a facial expression recognition system built upon the CLIP architecture. The authors evaluate CLIPER against state-of-the-art static facial expression recognition (SFER) and dynamic facial expression recognition (DFER) methods on widely used benchmarks. Their results demonstrate that CLIPER outperforms existing approaches in terms of recognition accuracy.

Lin et al. [2024] propose a lightweight framework capable of efficiently addressing both facial expression classification and Action Unit (AU) detection. The framework leverages a frozen CLIP image encoder to extract visual features. Experimental results on the Aff-Wild2 dataset [Kollias and Zafeiriou, 2019] show that it outperforms the baseline model—based on a VGG16 architecture [Simonyan and Zisserman, 2015] with fixed convolutional layers—while maintaining low computational overhead.

Collectively, these studies highlight the emerging potential of large language models and foundation models in biometric applications, underscoring their versatility, accuracy, and capacity for interpretability.

3 Methodology

We want to evaluate the capabilities of VLMs on a diverse set of soft and hard biometric tasks, including pairwise verification (e.g., face, fingerprint, iris), binary classification (e.g., gender recognition, attribute prediction), and age estimation. Each task is formulated as a natural language prompt provided to the VLM, which returns a direct answer. Our methodology involves prompt-based inference followed by log-probability analysis to determine the model's confidence and select its prediction accordingly.

Binary classification and verification tasks. For classification and verification tasks where the expected answers are discrete—typically "yes" or "no"—we begin by constructing a task-specific prompt (e.g., *"Do these two fingerprint images belong to the same finger? Answer with only 'yes' or 'no'."*) and feed the prompt along with the input image(s) to the VLM. The model then generates a one-token response. We extract the log probabilities of the first generated token corresponding to each possible answer and we normalized them using the softmax function. The predicted label is the one with the higher normalized probability. This technique allows us to make deterministic predictions while also evaluating the model's confidence.

Age estimation task. To evaluate age estimation capabilities, we prompt the VLM with a natural-language instruction and ask it to directly predict the person's age from an image. Unlike binary classification tasks, the model must output a discrete numerical value representing age, which we interpret through a candidate scoring strategy based on output token log-probabilities. For each image, we construct a dialogue-like message where the user uploads the photo and asks a prompt such as *"How old is the person in this image? Answer with a single number."* The model is instructed to respond with a single integer, which we constrain to a plausible age range $[a_{\min}, a_{\max}]$ (e.g., 0 to 101 years). We systematically evaluate all possible candidate age values within the predefined range. For each age $a \in [a_{\min}, a_{\max}]$, we compute the log-probability of the model generating the exact age string (e.g., "42") followed by an end-of-turn token. After scoring all candidates, we apply a softmax over the resulting log-probabilities to obtain a probability distribution over age values, and select the age with the highest probability as the model's prediction.

4 Experiments

In this section, we present a series of experiments aimed at evaluating our models across various biometric and soft biometric tasks. We begin by outlining the specific tasks and their corresponding benchmark datasets. Next, we describe the models used for evaluation. Finally, we detail the metrics employed to assess the models' performance on each task.

4.1 Tasks and datasets

We selected five biometric tasks, encompassing both soft and hard biometrics, and chose benchmark datasets to evaluate their performance.

4.1.1 Face Verification

Face recognition involves determining the identity of an individual from a facial image. The term "face recognition" can encompass several tasks, one of which is pairwise matching. This task requires verifying whether two facial images depict the same individual. We evaluate this task using the *Labeled Faces in the Wild* (LFW) [Huang et al., 2008] dataset, which contains images captured in unconstrained real-world conditions with variations in pose, lighting, background, and subject appearance. The benchmark provides 6000 predefined image pairs split evenly between matched (same person) and mismatched (different people) pairs.

To assess model performance, we employed a prompt in which two face images were provided simultaneously, and the model was asked: *"Do these two images show the same person? Answer with only 'yes' or 'no'."*

4.1.2 Age and Gender Estimation

Age and gender estimation are soft biometric tasks focused on predicting an individual's age or gender based on facial features. Unlike strong biometrics, such as fingerprints or iris patterns, these attributes provide probabilistic cues rather than unique identifiers.

We use the *AgeDB* dataset [Moschoglou et al., 2017], which includes 16488 facial images of 568 well-known individuals annotated with identity, age, and gender. The dataset spans a wide age range (1 to 101 years), with an average of 29 images per person and an average intra-subject age range of 50.3 years.

For age estimation, we prompted the model with each image and asked: *"How old is the person in this image? Answer with only a number representing their age."* For gender classification, we used: *"What is the gender of the person in this image? Answer with only 'male' or 'female'."*

4.1.3 Iris Verification

Iris recognition is a biometric identification technique that uses the unique texture patterns of the iris—the colored part of the eye surrounding the pupil. Due to its stability and high distinctiveness, iris recognition is considered highly accurate for verifying individual identities.

We used the *CASIA-Iris-Thousand* dataset (CASIA from now on), which includes 20000 iris images from 1000 subjects. For evaluation, we created 20000 image pairs, evenly split between genuine (same subject) and impostor (different subjects) pairs, with eye selection randomized.

The prompt presented two iris images and asked the model: *"Do these two iris images belong to the same person? Answer with only 'yes' or 'no'."*

4.1.4 Fingerprint Verification

Fingerprint verification is the task of determining whether two fingerprint images originate from the same individual. This is typically done by analyzing ridge patterns and minutiae points and is considered a strong biometric modality.

We evaluated models using data from the Fingerprint Verification Competitions (FVC) from 2000, 2002, and 2004 [Maltoni et al., 2009]. Each competition includes four databases, totaling 12 datasets. In each dataset, the images represent fingerprints from 110 individuals with 8 samples each. The datasest includes 3080 genuine pairs and 4995 impostor pairs.

The model was prompted with: *"Do these two fingerprint images belong to the same finger? Answer with only 'yes' or 'no'."*

4.1.5 Face Attribute Classification

Face attribute prediction involves identifying binary facial attributes such as age group, presence of facial hair, eyewear, expression, or gender. These soft biometric attributes offer valuable information for tasks such as demographic analysis, identity inference, and image tagging.

We used the evaluation split of the *CelebA* dataset [Liu et al., 2015], which includes 202600 images labeled with 40 binary facial attributes. Each image was input to the model along with a specific attribute prompt.

For each attribute, the model was asked: *"Does the person in the image have the attribute 'X'? Answer with only 'yes' or 'no'."* where X represents a specific attribute (e.g., "Smiling," "Wearing Hat," "Male").

4.2 Models

We evaluate three versions of the recently released Gemma 3 models [Gemma Team, 2025], a family of lightweight, open-source, multimodal LLMs developed by Google and Google DeepMind. Gemma 3 extends the previous Gemma family by introducing vision understanding capabilities, longer context windows (up to 128K tokens), and enhanced multilingual and instruction-following performance. Specifically, we use the instruction-tuned variants available on Hugging Face: `google/gemma-3-4b-it`, `google/gemma-3-12b-it`, and `google/gemma-3-27b-it`, comprising 4, 12, and 27 billion parameters, respectively.

4.3 Evaluation Metrics

Verification tasks. In the context of verification tasks—such as face, iris, and fingerprint recognition—we assess the performance of binary classification systems based on varying decision thresholds τ. The effectiveness of these systems is evaluated using several well-established metrics.

The *False Acceptance Rate (FAR)* quantifies the proportion of impostor pairs (i.e., image pairs from different individuals) that are incorrectly classified as genuine. Conversely, the *False Rejection Rate (FRR)* captures the proportion of genuine pairs (i.e., image pairs from the same individual) that are mistakenly classified as impostors. Both metrics depend on the classification threshold τ, with FRR specifically being equal to $1-\text{TPR}$, where TPR is the True Positive Rate, and FAR being equal to the False Positive Rate (FPR).

To provide a concise and threshold-independent performance summary, we compute the *Equal Error Rate (EER)*—the point at which FAR and FRR are equal (or as close as possible). A lower EER indicates a more accurate verification system. We also report the classification *accuracy*, FAR, and FRR at the EER threshold τ_{EER}, offering insights into performance when the system is tuned for equal false acceptance and rejection trade-offs.

Another threshold-independent metric is the *Area Under the ROC Curve (AUC)*. The ROC curve plots TPR against FPR at various thresholds, and the AUC measures the system's overall ability to distinguish between genuine and impostor pairs. An AUC score closer to 1 indicates superior discriminatory power.

Binary classification tasks. For binary gender classification and face attribute classification tasks (i.e., predicting "male" or "female"), we evaluate performance using standard classification metrics. These include *accuracy*, *precision*, *recall*, and *F1-score* for each class, alongside the *confusion matrix*, which illustrates the distribution of correct and incorrect predictions across the classes.

Age estimation task. Age estimation, a regression-based task, is evaluated using different criteria. The primary metric is the *Mean Absolute Error (MAE)*, which measures the average magnitude of prediction errors, regardless of direction. To complement MAE, we compute the *Cumulative Score (CS)*, which reports the percentage of predictions whose absolute error falls within a specified range (e.g., 1, 3, 5, 7, or 10 years). This score reflects how often the model makes acceptably close predictions. Additionally, we report the *standard deviation of the errors*, which captures the consistency of the model's predictions across the dataset.

5 Results

5.1 Face Verification on Labeled Faces in the Wild

We evaluate face verification performance using the LFW dataset [Huang et al., 2008]. As shown in Table 1, all three Gemma 3 models achieve strong performance in the zero-shot setting, with accuracy improving as model size increases. The largest model, Gemma 3 27B, reaches 95.93% accuracy and an EER of 4.07%, outperforming its smaller variants. Although the SOTA model—ArcFace [Deng et al., 2022]—achieves 99.83% accuracy, it is specifically trained for this task using deep convolutional networks and preprocessed features, while Gemma 3 operates in a zero-shot setting on a raw pair of images.

Model	EER↓	AUC↑	Acc↑	FAR↓	FRR↓
Gemma 3 4B	8.55	0.97	91.45	8.53	8.57
Gemma 3 12B	<u>4.63</u>	**0.99**	95.37	<u>4.60</u>	<u>4.67</u>
Gemma 3 27B	**4.07**	**0.99**	<u>95.93</u>	**4.00**	**4.13**
SOTA [Deng et al., 2022]	-	-	**99.83**	-	-

Table 1: Performance of Gemma 3 models and SOTA model on face verification (Labeled Faces in the Wild [Huang et al., 2008]). EER is expressed in %; accuracy, FAR, and FRR are computed using the threshold selected at the EER threshold.

The ROC curves in Figure 1 illustrate the trade-off between the true positive rate and false positive rate at various thresholds. The curve for the 27B model closely approaches the top-left corner, indicating high

discriminative ability. Similarly, the Detection Error Tradeoff (DET) curves in Figure 2 show that larger models maintain lower false rejection and false acceptance rates. Curves closer to the bottom-left indicate better performance, and again, the 27B model demonstrates the most favorable trade-offs.

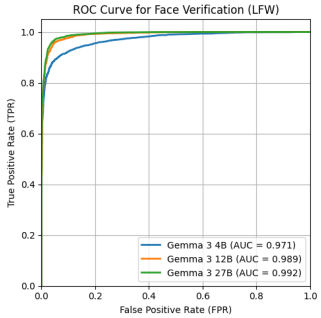

Figure 1: ROC Curves for each Gemma 3 model for face verification task on LFW.

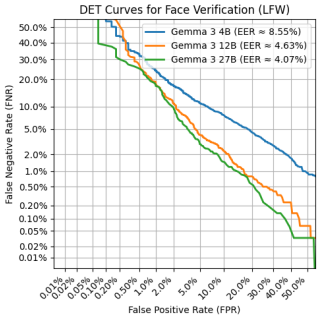

Figure 2: Detection Error Trade-off Curves for each Gemma 3 model for face verification task on LFW.

5.2 Age estimation on AgeDB

Gemma models struggle with the AgeDB age estimation task. The best-performing model, Gemma 3 12B, achieves a MAE of 13.02 years—substantially higher than the current SOTA MAE of 5.55 achieved by Kuprashevich and Tolstykh [2023], who propose a specialized vision transformer-based model. Moreover, all three Gemma variants exhibit high standard deviations in prediction error, indicating unstable and inconsistent estimates across samples.

This is further corroborated by the low cumulative score values: fewer than half of the predictions fall within 10 years of the ground truth for any of the models, with Gemma 3 12B reaching only 47.65% at CS@10yr.

Model	**MAE↓ (Years)**	**Error↓ Std. Dev.**	**Cumulative Score within X years↑ (%)**				
			1yr	**3yr**	**5yr**	**7yr**	**10yr**
Gemma 3 4B	21.36	17.03	4.43	10.46	16.19	21.80	30.12
Gemma 3 12B	13.02	**12.19**	**7.69**	**17.32**	**26.58**	**35.36**	**47.65**
Gemma 3 27B	19.24	21.69	5.40	12.95	19.32	25.81	35.21
SOTA [Kuprashevich and Tolstykh, 2023]	**5.55**	-	-	-	-	-	-

Table 2: Age estimation performance of Gemma 3 models on AgeDB. MAE and standard deviation (Std. Dev.) are reported in years. CS@Xyr refers to the percentage of predictions with absolute error less than or equal to X years. The SOTA results are reported from Kuprashevich and Tolstykh [2023].

Figure 3 illustrates the cumulative score curves across varying error tolerances. Steeper curves represent stronger performance. Notably, Gemma 3 12B consistently outperforms the smaller and larger variants, suggesting that model size alone does not directly correlate with better age estimation in this setting.

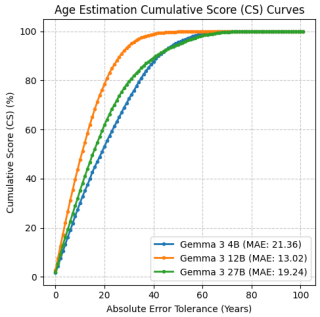

Figure 3: Cumulative Score (CS) vs. Absolute Error Tolerance on AgeDB. Steeper curves indicate better model performance, as a greater proportion of predictions fall within smaller error margins.

5.3 Gender Classification on AgeDB

In the gender classification task, zero-shot performance from the Gemma 3 models is remarkably close to that of SOTA methods. The best-performing Gemma model, Gemma 3 27B, achieves an accuracy of 98.00%, nearly matching the 98.3% accuracy reported by Kuprashevich and Tolstykh [2023]. Model size does not show a strict correlation with accuracy: the smaller Gemma 3 4B slightly outperforms the 12B variant, achieving 97.99% accuracy, compared to 97.64%. These results suggest that even relatively lightweight large language models can achieve robust performance in zero-shot gender classification.

Model	Overall	Female			Male		
	Accuracy (%)	**F1**	**Precision**	**Recall**	**F1**	**Precision**	**Recall**
Gemma 3 4B	97.99	**0.98**	**0.97**	**0.98**	**0.98**	**0.99**	**0.98**
Gemma 3 12B	97.64	0.97	0.97	**0.97**	**0.98**	0.98	**0.98**
Gemma 3 27B	<u>98.00</u>	**0.98**	**0.97**	**0.98**	**0.98**	**0.99**	**0.98**
SOTA	**98.3**	-	-	-	-	-	-

Table 3: Performance of Gemma 3 models and the SOTA model [Kuprashevich and Tolstykh, 2023] on gender classification. Reported metrics include overall accuracy, as well as per-class precision, recall, and F1-score for female and male categories.

Confusion matrices (Figure 4) provide further support for these results. They show a high rate of true positives and true negatives, with minimal false positives and false negatives. This confirms that the models are making reliable predictions across both gender classes.

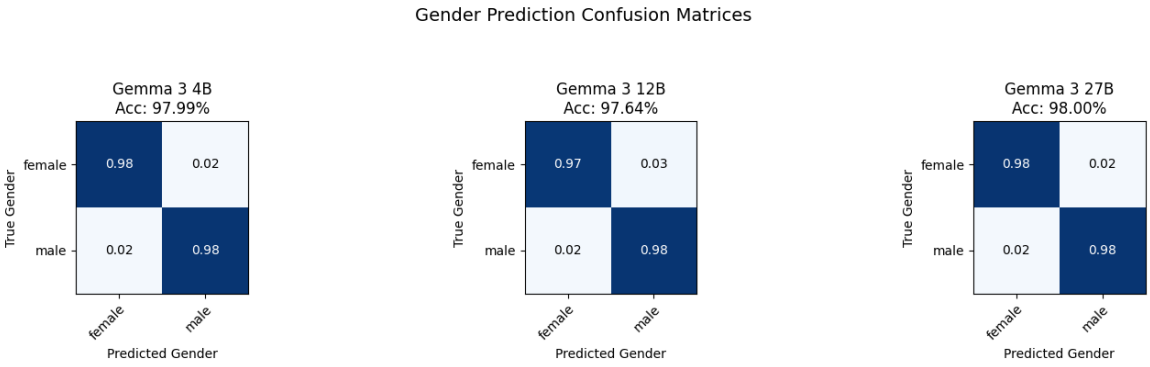

Figure 4: Confusion matrix for gender classification using Gemma 3 models.

5.4 Iris Verification on CASIA-Iris-Thousand

The SOTA accuracy on the CASIA-Iris-Thousand dataset is 99.68%, achieved by a specialized system proposed in Kotsuwan et al. [2025], which integrates iris and periocular verification using support vector classification. In contrast, our best-performing zero-shot model, Gemma 3 27B, achieves a verification accuracy of 70.06% (see Table 4). Although significantly lower than the SOTA result, this outcome is still promising considering the lack of fine-tuning. Moreover, results suggest that model performance improves with scale, as larger models show lower EER and higher AUC.

Model	**EER (%)↓**	**AUC↑**	**Accuracy (%)↑**	**FAR (%)↓**	**FRR (%)↓**
Gemma 3 4B	35.05	0.70	64.94	35.00	35.11
Gemma 3 12B	<u>33.28</u>	<u>0.73</u>	66.72	<u>33.24</u>	<u>33.33</u>
Gemma 3 27B	**29.94**	**0.76**	<u>70.06</u>	**30.15**	**29.73**
SOTA [Kotsuwan et al., 2025]	-	-	**99.68**	-	-

Table 4: Performance of Gemma 3 models and a SOTA model [Kotsuwan et al., 2025] on iris verification using the CASIA-Iris-Thousand dataset. Accuracy, FAR, and FRR are reported at the EER threshold.

The ROC curves in Figure 5 illustrate moderate discriminative ability for the Gemma models, though far from ideal. As commented previously, performance improves steadily with model size. The DET curves (Figure 6) show a balanced trade-off between false acceptance and false rejection rates. Despite high error rates, the linear relationship indicates no strong bias toward either type of error, suggesting the models are not skewed toward false positives or false negatives.

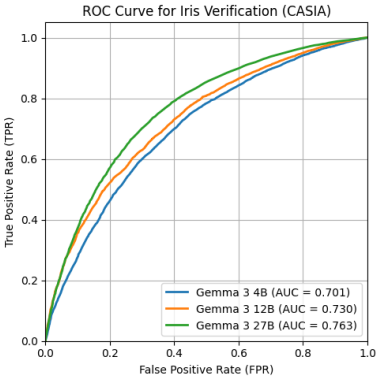

Figure 5: ROC curves for Gemma 3 models on iris verification with CASIA-Iris-Thousand. Larger models demonstrate improved separation between positive and negative pairs.

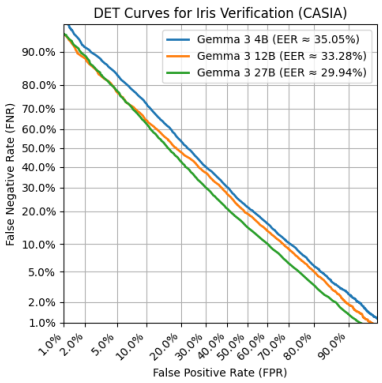

Figure 6: DET curves for Gemma 3 models on CASIA. Balanced but high error rates highlight the difficulty of zero-shot iris verification.

5.5 Fingerprint Verification on FVC Datasets

The Gemma 3 models show limited capability in fingerprint verification. The best-performing model, Gemma 3 12B, achieves a mean EER of 60.37%, which is far from competitive with current SOTA systems (see Table 6). For reference, the leading fingerprint verification system—developed by a private company—reports an EER of just 0.014%.

These high-performing systems typically rely on deep learning-based feature extractors specifically tailored to the fingerprint modality. Furthermore, fingerprint verification is a mature field where traditional algorithms already achieve high accuracy using handcrafted features, potentially reducing the value added by general-purpose models in a zero-shot setting.

Model	FVC2000 Acc.↑	FVC2002 Acc.↑	FVC2004 Acc.↑	Mean EER (%)↓
Gemma 3 4B	32.00	36.57	37.16	64.76
Gemma 3 12B	**34.90**	**41.41**	**42.56**	60.37
Gemma 3 27B	30.07	38.23	39.80	63.98
SOTA (Private)	-	-	-	**0.014**

Table 5: Accuracy and mean Equal Error Rate (EER) of Gemma 3 models on FVC2000, FVC2002, and FVC2004 fingerprint verification datasets.

5.6 Face Attribute detection on CelebA

We compare our results with the strongest baseline reported in Liu et al. [2015], which uses a CNN architecture similar to AlexNet (LNets+ANet). While all Gemma 3 models underperform relative to this baseline, their performance remains within a comparable range across many attributes.

Among the facial attributes, the Gemma 3 models perform best on Male and Bald, which are also among the highest-performing attributes for the SOTA baseline. Conversely, the Gemma models struggle most with Narrow Eyes and Oval Face, while the SOTA model shows its lowest performance on Oval Face and Big Lips. Tables 6 and 7 show all results on this task .

Model	5 o Clock Shadow	Arched Eyebrows	Attractive	Bags Under Eyes	Bald	Bangs	Big Lips	Big Nose	Black Hair	Blond Hair	Blurry	Brown Hair	Bushy Eyebrows	Chubby	Double Chin	Eyeglasses	Goatee	Gray Hair	Heavy Makeup	High Cheekbones	Male
Gemma 3 4B	80.99	55.12	73.39	47.18	93.38	82.85	58.21	62.02	80.46	90.70	73.53	66.54	61.09	78.08	77.18	84.37	90.90	91.25	77.82	50.87	98.48
Gemma 3 12B	81.37	65.47	74.42	53.77	94.56	86.56	55.61	61.05	83.11	91.65	76.64	70.73	56.34	73.99	72.72	95.83	88.79	90.72	69.80	50.89	**98.60**
Gemma 3 27B	85.00	71.80	76.15	62.23	93.85	87.04	60.05	69.37	83.75	91.78	79.46	72.66	60.69	81.03	82.04	95.31	92.27	89.77	77.22	57.41	98.49
SOTA Liu et al. [2015]	**91.00**	**79.00**	**81.00**	**79.00**	**98.00**	**95.00**	**68.00**	**78.00**	**88.00**	**95.00**	**84.00**	**80.00**	**90.00**	**91.00**	**92.00**	**99.00**	**95.00**	**97.00**	**90.00**	**87.00**	98.00

Table 6: Facial attribute classification accuracy at EER (%) – Part 1.

Model	Mouth Slightly Open	Mustache	Narrow Eyes	No Beard	Oval Face	Pale Skin	Pointy Nose	Receding Hairline	Rosy Cheeks	Sideburns	Smiling	Straight Hair	Wavy Hair	Wearing Earrings	Wearing Hat	Wearing Lipstick	Wearing Necklace	Wearing Necktie	Young	Average
Gemma 3 4B	70.30	87.20	48.15	88.32	54.85	65.09	43.28	58.57	71.31	87.25	89.85	65.52	71.82	85.52	96.36	88.37	67.81	90.13	83.10	76.07
Gemma 3 12B	71.44	87.28	47.61	91.01	54.89	65.83	52.64	61.32	68.03	68.87	90.32	70.75	79.28	85.31	96.46	84.82	67.30	89.53	82.76	75.56
Gemma 3 27B	81.90	89.75	47.03	90.35	56.43	67.07	48.39	60.75	70.77	85.80	90.67	70.70	76.57	**85.91**	96.20	89.61	67.32	90.60	83.56	76.94
SOTA Liu et al. [2015]	**92.00**	**95.00**	**81.00**	**95.00**	**66.00**	**91.00**	**72.00**	**89.00**	**90.00**	**96.00**	**92.00**	**73.00**	**80.00**	82.00	**99.00**	**93.00**	**71.00**	**93.00**	**87.00**	**87.00**

Table 7: Facial attribute classification accuracy at EER (%) – Part 2.

6 Conclusions

This report has undertaken an evaluation of the Gemma 3 family of VLMs across a diverse spectrum of biometric tasks, operating in a zero-shot, prompt-based inference setting. Our analysis indicate a varied landscape of capabilities, highlighting both promising avenues and significant challenges for current VLMs in the biometrics domain. The Gemma 3 models, particularly the larger 27B variant, demonstrated strong zero-shot performance in face verification on the LFW dataset and remarkably high accuracy in gender classification on AgeDB, nearing SOTA levels achieved by specialized, trained models. Performance in iris verification, while considerably lower than SOTA, showed improvement with model scale and indicated potential for zero-shot applicability.

Conversely, the models struggled significantly with age estimation, exhibiting high MEA and inconsistent predictions, falling far short of specialized vision transformer-based approaches. Similarly, fingerprint verification proved to be a challenging modality, with performance levels significantly below those of established systems that leverage highly specialized feature extractors. For face attribute classification on CelebA, while the Gemma 3 models did not match the SOTA CNN-based baseline, their performance was comparable across many attributes, particularly for more distinct features like 'Male' and 'Bald'.

Overall, the Gemma 3 models showcase the emerging potential of VLMs to handle certain biometric tasks effectively without task-specific training. However, for modalities requiring fine-grained detail analysis like fingerprints, or complex regression tasks like precise age estimation, current general-purpose VLMs appear less suited in a zero-shot context. The results also suggest that while model scale often correlates with improved performance, this is not universally true across all tasks, as seen in age estimation where the 12B model outperformed the 27B variant.

References

T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. Language models are few-shot learners. In H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

I. Deandres-Tame, R. Tolosana, R. Vera-Rodriguez, A. Morales, J. Fierrez, and J. Ortega-Garcia. How good is chatgpt at face biometrics? a first look into recognition, soft biometrics, and explainability. *IEEE Access*, 12:34390–34401, 2024. doi: 10.1109/ACCESS.2024.3370437.

J. Deng, J. Guo, J. Yang, N. Xue, I. Kotsia, and S. Zafeiriou. Arcface: Additive angular margin loss for deep face recognition. *IEEE Trans. Pattern Anal. Mach. Intell.*, 44(10_Part_1):5962–5979, Oct. 2022. ISSN 0162-8828. doi: 10.1109/TPAMI.2021.3087709. URL `https://doi.org/10.1109/TPAMI.2021.3087709`.

J. Devlin, M. Chang, K. Lee, and K. Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In J. Burstein, C. Doran, and T. Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics, 2019. doi: 10.18653/V1/N19-1423. URL `https://doi.org/10.18653/v1/n19-1423`.

A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net, 2021. URL `https://openreview.net/forum?id=YicbFdNTTy`.

P. Farmanifard and A. Ross. Chatgpt meets iris biometrics. In *2024 IEEE International Joint Conference on Biometrics (IJCB)*, pages 1–10, 2024. doi: 10.1109/IJCB62174.2024.10744525.

Gemini Team. Gemini: A family of highly capable multimodal models, 2025. URL https://arxiv.org/abs/2312.11805.

Gemma Team. Gemma 3 technical report, 2025. URL https://arxiv.org/abs/2503.19786.

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, 2024. doi: 10.1109/ICIP51287.2024.10647924.

G. B. Huang, M. Mattar, T. Berg, and E. 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*, Marseille, France, Oct. 2008. Erik Learned-Miller and Andras Ferencz and Frédéric Jurie. URL https://inria.hal.science/inria-00321923.

K. Karkkainen and J. Joo. 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*, pages 1548–1558, 2021.

M. Kim, A. K. Jain, and X. Liu. Adaface: Quality adaptive margin for face recognition. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 2022.

D. Kollias and S. Zafeiriou. Expression, affect, action unit recognition: Aff-wild2, multi-task learning and arcface, 2019. URL https://arxiv.org/abs/1910.04855.

O. Kotsuwan, C. Chokchaisiri, W. Kongprawechnon, S. Duangpummet, K. Galajit, and J. Karnjana. Enhance biometric authentication: Integrating iris and periocular verification through support vector classification. In *Integrated Uncertainty in Knowledge Modelling and Decision Making: 11th International Symposium, IUKM 2025, Ho Chi Minh City, Vietnam, March 17–19, 2025, Proceedings, Part I*, page 163–173, Berlin, Heidelberg, 2025. Springer-Verlag. ISBN 978-981-96-4605-0.

M. Kuprashevich and I. Tolstykh. Mivolo: Multi-input transformer for age and gender estimation, 2023. URL https://arxiv.org/abs/2307.04616.

H. Li, H. Niu, Z. Zhu, and F. Zhao. Cliper: A unified vision-language framework for in-the-wild facial expression recognition. In *2024 IEEE International Conference on Multimedia and Expo (ICME)*, pages 1–6, 2024. doi: 10.1109/ICME57554.2024.10687508.

L. Lin, S. Papabathini, X. Wang, and S. Hu. Robust light-weight facial affective behavior recognition with clip. In *2024 IEEE 7th International Conference on Multimedia Information Processing and Retrieval (MIPR)*, pages 607–611, 2024. doi: 10.1109/MIPR62202.2024.00103.

Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *2015 IEEE International Conference on Computer Vision, ICCV 2015, Santiago, Chile, December 7-13, 2015*, pages 3730–3738. IEEE Computer Society, 2015. doi: 10.1109/ICCV.2015.425.

D. Maltoni, D. Maio, A. K. Jain, S. Prabhakar, et al. *Handbook of fingerprint recognition*, volume 2. Springer, 2009.

S. Moschoglou, A. Papaioannou, C. Sagonas, J. Deng, I. Kotsia, and S. Zafeiriou. Agedb: The first manually collected, in-the-wild age database. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, July 2017.

Neurotechnology. Verieye sdk, 2025. URL https://www.neurotechnology.com/verieye.html.

OpenAI. Gpt-4v(ision) system card, 2023. URL https://cdn.openai.com/papers/GPTV_System_Card.pdf.

A. Radford, J. W. Kim, C. Hallacy, A. Ramesh, G. Goh, S. Agarwal, G. Sastry, A. Askell, P. Mishkin, J. Clark, G. Krueger, and I. Sutskever. Learning transferable visual models from natural language supervision. In M. Meila and T. Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 8748–8763. PMLR, 2021. URL http://proceedings.mlr.press/v139/radford21a.html.

H. O. Shahreza and S. Marcel. Foundation models and biometrics: A survey and outlook. *TechRxiv*, 2025. doi: 10.36227/techrxiv.174119169.94570936/v1.

K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition, 2015. URL https://arxiv.org/abs/1409.1556.