Exploring Biometrics Through the Lens of Generative AI and Large Vision-Language Models

Vishal M. Patel

Associate Professor

Department of Electrical and Computer Engineering



vpatel36@jhu.edu

https://engineering.jhu.edu/vpatel36/

January 14, 2025

IAPR/IEEE WINTER SCHOOL ON BIOMETRICS 2025

12-16 January 2025 Shenzhen, China

History of Artificial Neural Networks



ImageNet Challenge

- Large Scale Visual Recognition Challenge (ILSVRC)
 - 1000 object categories
 - 1.2M training images



Classification Results (CLS)



Russakovsky et al. ILSVRC2017 CVPRW 2017

Face Recognition

- Labeled Faces in the Wild (LFW)
 - 5,749 subjects
 - 13,233 faces



 Mean classification accuracies: YI+AI (0.9983 ± 0.0024)
 FRDC (0.9972 ± 0.0029)
 CHTFace (0.9960 ± 0.0025)



Training data: 4 mil faces, 4000 identities (Facebook)

Taigman et al. DeepFace: Closing the Gap to Human-Level Performance in Face Verification, CVPR 2014 .

Segment Anything Model (SAM)

- A new AI model from Meta AI that can "cut out" any object, in any image, with a single click
 - Trained using 1.1 billion segmentation masks from 11 million images





https://segment-anything.com/

LeNet5 vs AlexNet



- Trained on MNIST digit dataset with 60K training examples
- Sigmoid or tanh nonlinearity
- Average pooling
- Fully connected layers at the end

AlexNet Krizhevsky et al. 2012



- Trained on ImageNet dataset with 1.2M training images
- Rectified Linear Unit (ReLU) nonlinearity
- Max pooling
- GPU implementation
 - Trained on two GPUs for a week
- Dropout regularization
- Fully connected layers at the end

Why?

- Availability of large annotated datasets
- More layers
 - Capture more invariances
- More compute
 - Availability and affordability of GPUs
- Better regularization
 - Dropout
- New nonlinearities
 - Rectified Linear Unit (ReLU)
 - Parametric Rectified Linear Unit (PReLU)



Recent Large Pre-Trained Models



https://alinakhay.com/Comprehensive-Guide-to-the-State-of-Large-Language-Models

Scaling Laws for Neural Language Models



Figure 1 Language modeling performance improves smoothly as we increase the model size, datasetset size, and amount of compute² used for training. For optimal performance all three factors must be scaled up in tandem. Empirical performance has a power-law relationship with each individual factor when not bottlenecked by the other two.

There is a power law relationship between the number of parameters in an autoregressive language model and its performance!

- Train larger models and expect performance improvements.
- Model size and the number of training tokens should be scaled in equal proportions.

Kaplan et al. arXiv 2020 Chinchilla, DeepMind 2022

Availability of large annotated data

- More layers
 - Capture more invariances
- More computing
 - Availability and affordability of GPUs
- Better regularization
 - Dropout
- New nonlinearities
 - Rectified Linear Unit (ReLU)
 - Parametric Rectified Linear Unit (PReLU)



Large Datasets

- Collecting and annotating datasets
 - Expensive
 - Labor intensive
 - User privacy issues
 - GDPR: General Data
 Protection Regulation
 - HIPAA: Health Insurance Portability and Accountability Act, 1996
 - SHIELD: Stop Hacks and Improve Electronic Data Security Act, Jan 1 2019
 - PCI: Payment Card Industry Data Security Standard, 2004
 - IRB: Institutional Review
 Board



Protecting User Privacy

- Data privacy (protect the data)
 - Cancelable biometrics
 - Modify data through revocable and noninvertible transformations
 - BioHashing
 - Random projections are used to generate templates
 - Differential privacy
 - An algorithm is differentially private if its behavior hardly changes when a single individual joins or leaves the dataset
 - Hide unique samples (add noise to data)
 - Homomorphic encryption
 - Perform calculations on encrypted data

V. M. Patel, N. K. Ratha and R. Chellappa, "Cancelable Biometrics: A review," in IEEE Signal Processing Magazine, vol. 32, no. 5, pp. 54-65, Sept. 2015.





Federated Learning

- Federated learning (build protection into the models)
 - Multiple entities collaboratively train a model while ensuring that their data remains decentralized.
 - Communication-efficient learning of deep networks from decentralized data, AISTATS 2017, McMahan et al. (Google)



Federated Learning - FedAvg



- Different users (clients) collaboratively learn a machine learning model with the help of a server
- Local training
 - Users locally compute training parameters and send them to the server
- Model aggregating
 - The server performs secure aggregation over the uploaded parameters from different users without learning local information
- Parameters broadcasting
 - The server broadcasts the aggregated parameters to the users
- Model updating
 - All users update their respective models with aggregated parameters and test the performance of the updated models

Federated Learning - Applications

- Next word prediction (Google)
 - Federated Learning for Mobile Keyboard
 Prediction, Hard et al., 2018
- Speaker recognition (Apple Siri)
 - QuickType (Apple's personalized keyboard)



Fig. 1. Next word predictions in Gboard. Based on the context "I love you", the keyboard predicts "and", "too", and "so much".



Artificial intelligence / Machine learning

How Apple personalizes Siri without hoovering up your data

The tech giant is using privacy-preserving machine learning to improve its voice assistant while keeping your data on your phone.

Synthetic Data Generation



Generative Models













Image Replication in Diffusion Models

- Diffusion models have outperformed GANs in image synthesis and have demonstrated a remarkable ability in generating high-quality images
- Despite their potential, diffusionbased image synthesis models are known to replicate training data and generate biased samples.





Perera & Patel, IJCB 2023 Somepalli et al. CVPR 2023

Replication in Video Diffusion Models



Replication in Video Diffusion Models

Frame by Familiar Frame: Understanding Replication in Video Diffusion Models

Aimon Rahman *, Malsha V. Perera *, and Vishal M. Patel Johns Hopkins University

* denotes equal contribution

{arahma30,jperera4,vpatel36}@jhu.edu



Figure 1. Diffusion-based video synthesis models can sometimes replicate training data by assembling memorized foreground and background elements. We demonstrate this trend across multiple diffusion models trained on diverse datasets. Such occurrences prompt inquiries regarding data memorization and the ownership of videos produced by diffusion methods. Bottom row: Videos sourced from the RaMViD [18], VIDM [28], and LVDM [14] project websites. Top row: The most similar counterparts from the training dataset.

ш

Replication in Medical Diffusion Models



ONE HUNDRED EXAMPLES OF GPT-4 MEMORIZING CONTENT FROM THE NEW YORK TIMES

EXAMPLE 48: POPE IN BREAK FROM DOCTRINE BACKS SAMESEX CIVIL UNIONS

https://www.nytimes.com/2020/10/21/world/europe/pope-francis-same-sex-civil-unions.html Copyright number: TX 8-926-133 Effective Date: 2020-12-14

Prompt (taken from article):

ROME — Pope Francis expressed support for same-sex civil unions in remarks revealed in a

Output from GPT-4:

documentary film that premiered on Wednesday, a significant break from his predecessors that staked out new ground for the church in its recognition of gay people.

The remarks, coming from the leader of the Roman Catholic Church, had the potential to shift debates about the legal status of same-sex couples in nations around the globe and unsettle bishops worried that the unions threaten what the church considers traditional marriage — between one man and one woman.

"What we have to create is a civil union law. That way they are legally covered," Francis said in the documentary, "Francesco," which premiered at the Rome Film Festival, reiterating his view that gay people are children of God. "I stood up for

Actual text from NYTimes:

documentary film that premiered on Wednesday, a significant break from his predecessors that staked out new ground for the church in its recognition of gay people.

The remarks, coming from the leader of the Roman Catholic Church, had the potential to shift debates about the legal status of same-sex couples in nations around the globe and unsettle bishops worried that the unions threaten what the church considers traditional marriage — between one man and one woman.

"What we have to create is a civil union law. That way they are legally covered," Francis said in the documentary, "Francesco," which debuted at the Rome Film Festival, reiterating his view that gay people are children of God. "I stood up for

60

Social Biases in Generative Models

 Generative models have become widely popular across various applications, but they also risk perpetuating biases, which can lead to significant social consequences.



Gender distribution in the FFHQ training data (left) and generated data (right) using diffusion model.

Biased towards female class



Gender distribution in the FairFace-BW training data (left) and generated data (right) using diffusion model.

Biased towards male class



Age distribution in the FFHQ training data (left) and generated data (right) using diffusion model.

Biased towards biased towards generating younger faces



Racial distribution in FairFace-BW training data (left) and generated data (right) using diffusion model.

Biased towards white racial class

Key Observations

- Diffusion-based face generation models exhibit bias even when trained on balanced datasets.
- When trained with balanced datasets of varying sizes, diffusion models demonstrate bias towards certain attribute classes, with the degree of bias being dependent on the specific dataset and its size.
 - Complexity of the training subset might have played a role in the variation of attribute distribution within diffusion models. Factors such as lighting conditions, hairstyle, and makeup added complexity to the dataset.



Gender Distribution of FairFace-BW training data subsets and generated data : a) gender distribution in training subset, results on GAN and diffusion models trained with b) 300 samples, c) 700 samples, d) 1000 samples, e) 5000 samples, f) 10000 samples

Mitigating Replication and Bias in Diffusionbased Image Synthesis Models

• The proposed method involves fine-tuning the model using rewards to guide it toward a more desirable output distribution, encouraging the generation of preferred samples and suppressing undesirable ones.

Our objective is to solve the following optimization problem

 $\max_{\theta} \mathbb{E}_{\mathbf{x}_0 \sim p_{\theta}(\cdot)} [r(\mathbf{x}_0)].$

our goal of fine-tuning the diffusion model to maximize rewards can be approximately achieved by minimizing the following loss function with minimal alterations to the standard diffusion training

$$L_{reward} = \mathbb{E} \big[r(\mathbf{x}_0) \| \epsilon - \epsilon_{\theta}(\mathbf{x}_t, t) \|^2 \big].$$

Rewards

Mitigation of data replication:

$$\begin{aligned} r_{rep}(\mathbf{x}_0) &= 1 - \max_{\mathbf{x}_d \in \mathcal{D}} (deg(f(\mathbf{x}_0), f(\mathbf{x}_d))) \\ \\ L_{rep} &= L_{reward} + \gamma \mathbb{E}_{\mathbf{x}_d \sim \mathcal{D}} \big[\| \epsilon - \epsilon_{\theta}(\mathbf{x}_t, t) \|^2 \big] \end{aligned}$$

Mitigation of social biases:

$$r_{bias}(\mathbf{x}_0) = m_d + (m_d - m_c)^y + (m_c - m_d)^{(1-y)}.$$

Mitigation Results

Mitigating gender bias- FFHQ dataset



Results corresponding to the mitigation of gender bias in diffusion models trained with subsets of the FFHQ dataset. Each plot present the gender distribution of the generated samples of the diffusion models: pre-trained and tuned using proposed method. Plots (a), (b), (c), and (d) correspond to diffusion models trained with the entire dataset, 300 samples, 700 samples, and 1000 samples, respectively.

Mitigating racial bias- FairFace dataset



Results corresponding to the mitigation of racial bias in diffusion models trained with subsets of the FairFace-2 dataset. Each plot presents the racial distribution of the generated samples of the diffusion models: pre-trained and tuned using proposed method. Plots (a), (b), (c), and (d) correspond to the diffusion models trained with the entire dataset, 300 samples, 700 samples, and 1000 samples, respectively.

Mitigation Results

Simultaneous Mitigation data replication and social bias - FairFace dataset



Results corresponding to the simultaneous mitigation of data replication and social biases in diffusion models trained with 700 samples of the FairFace-2 dataset. Plots (a), (b), and (c) present the results for mitigation of data replication, gender bias, and racial bias, respectively.

Qualitative Results



LLMs for Face Understanding

• MLLMs have shown impressive abilities in a lot of domains but fail to answer basic questions related to face understanding.



FaceXBench

 5000 MCQ questions: 6 broad categories, 14 tasks, 25 public datasets

Statistic	Number
Total questions	5000
Total categories	6
Total tasks	14
Public datasets used	25
New dataset proposed (FaceXAPI)	1
Questions with multiple images	2750 (55%)
Questions with single image	2150 (43%)
Questions with only text	100 (2%)
Total images in all questions	11266
Unique number of images	10441
Unique question templates	757
Maximum question length	676
Maximum option length	207
Average question length	64.34
Average option length	11.04
Total options in each question	4
Frequency of A as correct option	1278 (25.56%)
Frequency of B as correct option	1332 (26.64%)
Frequency of C as correct option	1189 (23.78%)
Frequency of D as correct option	1201 (24.02%)



FaceXBench (6 categories and 14 tasks)

1. Bias and Fairness:

Age Estimation, Gender Prediction, Race Estimation

2. Face Recognition:

HR Face Recognition, LR Face Recognition, Celebrity identification

3. Face Authentication:

Face Anti-spoofing, Deepfake Detection

4. Face Analysis:

Attributes Prediction, Facial Expression Recognition

5. Face Localization:

Headpose Estimation, Face Parsing, Crowd Counting

6. Face Tools:

Face Tools Retrieval

Dataset	Number of Questions	Multiple Images	Single Images	Text Only	Task	Category
FairFace	300	200	100	0	Age Estimation	Bias & Fairness
UTKFace	200	150	50	0	Age Estimation	Bias & Fairness
FairFace	300	200	100	0	Gender Prediction	Bias & Fairness
UTKFace	200	150	50	0	Gender Prediction	Bias & Fairness
FairFace	300	200	100	0	Race Estimation	Bias & Fairness
UTKFace	200	150	50	0	Race Estimation	Bias & Fairness
LFW	60	60	0	0	HR Face Recognition	Face Recognition
AgeDB	100	100	0	0	HR Face Recognition	Face Recognition
CFP-FF	60	60	0	0	HR Face Recognition	Face Recognition
CFP-FP	60	60	0	0	HR Face Recognition	Face Recognition
CALFW	60	60	0	0	HR Face Recognition	Face Recognition
CPLFW	60	60	0	0	HR Face Recognition	Face Recognition
TinyFace	100	100	0	0	LR Face Recognition	Face Recognition
IMDB	300	150	150	0	Celebrity Identification	Face Recognition
WMCA	250	100	150	0	Face Anti-spoofing	Face Authentication
MSU-MFSD	50	50	0	0	Face Anti-spoofing	Face Authentication
CASIA-MFSD	50	50	0	0	Face Anti-spoofing	Face Authentication
ReplayAttack	50	50	0	0	Face Anti-spoofing	Face Authentication
CelebDF	150	150	0	0	Deepfake Detection	Face Authentication
FF++	150	150	0	0	Deepfake Detection	Face Authentication
CelebA	400	200	200	0	Attributes Prediction	Face Analysis
RAF-DB	200	100	100	0	Facial Expression Recognition	Face Analysis
AffectNet	200	100	100	0	Facial Expression Recognition	Face Analysis
AFLW2000	200	50	150	0	Headpose Estimation	Face Analysis
BIWI	200	50	150	0	Headpose Estimation	Face Analysis
JHUCrowd++	200	0	200	0	Crowd Counting	Face Localization
ShanghaiTech	100	0	100	0	Crowd Counting	Face Localization
CelebAMask-HQ	200	0	200	0	Face Parsing	Face Localization
LaPa	200	0	200	0	Face Parsing	Face Localization
FaceXAPI	100	0	0	100	Face Tools Retrieval	Face Tools Use

Age Estimation	Gender Prediction	Race Estimation	HR Face Recognition	LR Face Recognition	
Estimate the age group of the person in this image ?	Identify the gender of the person in the image ?	Which image shows a person belonging to the white race ?	Which two images are of the same person ?	The first image is of person A. Which of the other images are not person A?	
 (A) None of the above (B) 20-29 (C) 10-19 (D) 30-39 	(A) Male (B) Both (C) Female (D) None of the above	 (A) Both Images (B) Image 1 (C) Image 2 (D) None of the above 	 (A) Img 1 & Img 2 (B) Img 2 & Img 3 (C) Img 3 & Img 4 (D) Img 1 & Img 4 	(A) Image 3 (B) All images (C) Image 2 (D) Image 4	
Celebrity Identification	Face Anti-spoofing		Deepfake Detection	Attributes Prediction	
Identify the celebrity in the image ?	How many images are bonafide samples ?		Which deepfakes belong to the same identity ?	Which of the following attributes is NOT present in the image ?	
 (A) Christina Pickles (B) Kathleen Noone (C) Ilan Hall (D) Kyra Zagorsky 	(A) 1 (B) 2 (C) 3 (D) 0	FaceXBench	(A) Img 1 & Img 4 (B) Img 2 & Img 3 (C) Img 1 & Img 2 (D) Img 3 & Img 4	(A) Bangs (B) Rosy Cheeks (C) Wearing Lipstick (D) Wavy Hair	
Facial Expression Recognition	Crowd Counting	Face Parsing	Headpose Estimation	Face Tools Retrieval	
Which facial expression is the person showing in the image ?	How many people are present in this image ?	Which region is segmented out with white color?	What is the yaw angle range of headpose estimation for the person in	In an airport security setting, the system verifies traveler identity with high- resolution face recognition and gender classification. It should detect spoofing and then perform deepfake detection if Raged. Skip deepfake checks if spoofing confidence is high. Which API call sequence is correct?	
(A) contempt (B) happy (C) surprise (D) fear	(A) 2 (B) 5 (C) 6 (D) 7	(A) hat (B) eyeglasses (C) necklace (D) earring	this image ? (A) 0 to 10 (B) -40 to -30 (C) -20 to -10 (D) -10 to 0	 (A) api_7-identify_high_res_face, api_2-classify_gender, api_4-spoof_confidence_score, api_4-detect_spoofing, api_5-detect_deepfake (B) api_4-detect_spoofing, api_4-spoof_confidence_score, api_5-detect_deepfake, api_2-classify_gender, api_7-extract_high_res_embedding (C) api_7-identify_high_res_face, api_2-classify_gender, api_4-detect_spoofing, api_4-spoof_confidence_score, api_5-detect_deepfake (D) api_4-detect_spoofing, api_4-spoof_confidence_score, api_5-detect_deepfake (D) api_4-detect_spoofing, api_4-spoof_confidence_score, api_5-detect_deepfake (D) api_4-detect_spoofing, api_4-spoof_confidence_score, api_2-classify_gender 	

Results

- Open source models such as Qwen2VL and InternVL2 perform better than proprietary models such as GPT-40 and GeminiPro-1.5
- The max. accuracy achieved is less than 60% highlighting the challenge proposed by the benchmark.

Models	Overall	Bias &	Face Recognition	Face	Face A palveis	Face Localization	Face Tools Use	
(28)	(5,000)	(800)	(1,500)	(1,100)	(800)	(700)	(100)	
Random Choice	25.10	24.73	26.88	22.71	24.75	25.64	30.00	
Frequent Choice	32.22	30.73	29.50	40.14	33.25	29.73	40.00	
Open source MLLMs (< 4B parameters)								
PaliGemma [9]	32.22	35.67	26.50	28.00	37.62	32.27	12.00	
LLaVA-OneVision-0.5b-OV [41]	34.00	34.93	28.12	30.29	44.62	32.91	20.00	
VILA 1.5-3b [46]	35.80	38.27	33.25	30.86	44.50	31.82	28.00	
	Ор	en source M	1LLMs (4B - 1	3B parameters)				
Chameleon-7b [78]	17.04	10.27	17.12	6.86	20.25	28.91	33.00	
Eagle-X4-8B-Plus [69]	31.44	25.00	23.12	30.00	35.62	43.64	37.00	
Idefics-9b-Instruct [34]	34.58	37.93	28.62	34.43	37.38	34.18	15.00	
LLaVA-v1.5-7b [49]	36.22	41.20	33.12	30.14	43.50	32.18	15.00	
Monkey-Chat [45]	37.40	39.00	31.50	26.00	44.00	41.73	40.00	
MiniCPM-Llama3-v2.5 [95]	40.70	45.80	29.88	32.86	52.38	40.45	15.00	
LLaVA-NeXT-Interleave-7b [42]	43.80	52.53	38.00	38.57	55.88	32.27	26.00	
LLaVA-OneVision-7b-SI [41]	44.32	50.73	32.75	29.86	52.25	47.27	46.00	
Idefics2-8b [35]	44.52	52.67	31.25	33.57	53.25	43.91	42.00	
Mantis-SIGLIP-8b [27]	44.60	56.13	45.12	36.86	48.00	31.64	37.00	
Phi-3.5-Vision [1]	45.16	52.47	50.12	40.00	51.00	31.64	34.00	
LLaVA-OneVision-7b-OV [41]	48.98	61.40	38.38	35.57	55.12	44.82	38.00	
Qwen2-VL-7b-Instruct [85]	51.58	57.47	57.88	34.00	57.50	47.09	38.00	
InternVL2-8b [12]	53.24	62.40	61.75	35.43	55.38	45.09	45.00	
	0	pen source	MLLMs (> 13)	B parameters)				
Idefics-80b-Instruct [34]	35.86	39.87	35.12	27.71	35.12	38.55	15.00	
LLaVA-v1.5-13b [49]	39.88	44.60	34.88	34.14	44.75	37.27	39.00	
VILA 1.5-13b [46]	40.00	45.07	40.00	28.43	49.25	34.18	35.00	
CogVLM2-19b [22]	40.46	43.13	33.88	35.71	45.62	41.91	29.00	
InternVL-Chat-v1.5 [12]	49.18	59.73	41.38	33.00	55.12	46.73	46.00	
VILA 1.5-40b [46]	55.48	64.00	57.63	33.14	60.50	54.36	39.00	
LLaVA-OneVision-72b-OV [41]	56.42	66.53	52.00	37.43	63.25	53.73	48.00	
InternVL2-76b [12]	57.80	69.53	66.62	36.14	62.00	47.18	46.00	
Qwen2-VL-72b-Instruct [85]	57.86	62.20	69.12	41.14	57.88	55.45	46.00	
Proprietary MLLMs								
GPT-40 [25]	50.50	46.93	55.62	40.00	62.25	50.36	44.00	
GeminiPro 1.5 [79]	56.96	67.40	70.00	35.00	58.13	46.36	57.00	

Analysis

- Tasks such as crowd counting, deepfake detection, and headpose estimation are challenging, with most models performing poorly.
- 2. From the segmentation results, we observe that models trained using dynamic resolution exhibit superior performance in face parsing.





- 1. The model performs poorly on questions involving multiple images.
- For models that share the same vision encoder but differ in their LLM backbone, we observe that performance improves with the size of the LLM.



Summary

- Generative AI and LLMs promise to be an active area of research
- Open problems
 - Mitigation of data replication in video-based diffusion models
 - VLMs are not able to reason in tasks related to face understanding
 - Better VLMs for face understanding are needed!
 - Understanding and mitigating bias in LLMs

Acknowledgments





Call for Papers: IEEE TBIOM Special Issue on "Generative AI and Large Vision-Language Models for Biometrics"

November 11, 2024 | Biometrics Council News, Publications, T-BIOM

We invite submissions for the IEEE Transactions on Biometrics, Behavior, and Identity Science (T-BIOM) Special Issue on Generative AI and Large Vision-Language Models for Biometrics.

CALL FOR PAPERS	

IEEE Transactions on Biometrics, Behavior, and Identity Science (T-BIOM) Special Issue on Generative AI and Large Vision-Language Models for Biometrics

Submission Deadline: 31 May 2025

Targeted Publication: Q1 2026

Paper submission: https://ieee.atyponrex.com/journal/tbiom

The topics of interest of the special issue include, but are not limited to:

- + Novel generative AI models for responsible synthesis of biometric data
- + Novel generative models for conditional data synthesis
- + Biometrics interpretability and explainability through large language-vision models
- + Few-shot learning from large language-vision models

+ Generative AI and LVMs for detecting attacks on biometrics systems			
+ Generative AI-based image restoration	*** Important Dates ***		
+ Information leakage of synthetic data			
+ Data factories and label generation for biometric models	Submission deadline:	31 May 2025	
+ Quality assessment of AI generated data	First round of reviews completed (first decision):	August 2025	
+ Synthetic data for data augmentation	Second round of routing completed	Actaban 2025	
+ Detection of generated AI contents	Second round of reviews completed	UCLOBER 2025	
+ Bias mitigation using synthetic data	Final papers due	December 2025	
+ LLMs and VLMs for biometrics	Publication date:	Q1 2026	
+ Watermarking AI generated content			

- + New synthetic datasets and performance benchmarks
- + Security and privacy issues regarding the use of generative AI methods for biometrics
- + Ethical considerations regarding the use of generative AI methods for biometrics
- + Parameter efficient fine-tuning of VLMs for biometrics applications

More Information,

VISION & IMAGE

Vision and Image Understanding (VIU) Lab @JHU

https://engineering.jhu.edu/vpatel36/

Thank You!